The geometry of domain-general performance monitoring representations in the human medial frontal cortex

Zhongzheng Fu[1,2], Danielle Beam[1], Jeffrey M. Chung[3], Chrystal M. Reed[3], Adam N. Mamelak[1], Ralph Adolphs[2,4], Ueli Rutishauser[1,3,4,5,*]

[1] Department of Neurosurgery, Cedars-Sinai Medical Center, Los Angeles, CA, USA
[2] Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA, USA
[3] Department of Neurology, Cedars-Sinai Medical Center, Los Angeles, CA, USA
[4] Division of Biology and Bioengineering, California Institute of Technology, Pasadena, CA, USA
[5] Center for Neural Science and Medicine, Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, CA, USA

*Correspondence: rutishauseru@csmc.edu

Abstract

Flexibly adapting behavior to achieve a desired goal depends on the ability to monitor one's own performance. A key open question is how performance monitoring can be both highly flexible to support multiple tasks and specialized to support specific tasks. We characterized performance monitoring representations by recording single neurons in the human medial frontal cortex (MFC). Subjects performed two tasks that involve three types of cognitive conflict. Neural population representations of conflict, error and control demand generalized across tasks and time while at the same time also encoding task specialization. This arose from a combination of single neurons whose responses were task-invariant and non-linearly mixed. Neurons encoding conflict ex-post served to iteratively update internal estimates of control demand as predicted by a Bayesian model. These findings reveal how the MFC representation of evaluative signals are both abstract and specific, suggesting a mechanism for computing and maintaining control demand estimates across trials and tasks.

Introduction

Successful goal-directed behavior in uncertain environments depends critically on continual evaluation of one's own performance (Ullsperger, 2017; Ullsperger et al., 2014). We constantly evaluate whether we made an error, experienced conflict, received reward, or responded fast or slow. Information about past and present performance is in turn used by various downstream processes for cognitive control, affective responses and autonomic homeostasis. The resulting behavioral and physiological adaptations encompass task-specific attentional modulation of perception (Danielmeier et al., 2011; Egner and Hirsch, 2005; King et al., 2010; Purcell and Kiani, 2016), estimation of control demand (Darlington et al., 2018; Jiang et al., 2015; Shenhav et al., 2013), global modulation of motor system readiness (Aron et al., 2007; Danielmeier et al., 2011; King et al., 2010; Murphy et al., 2016; Niv et al., 2007; Wessel and Aron, 2017), emotional state (Bach and Dayan, 2017; Eldar et al., 2016; Shackman et al., 2011), or arousal levels (Crone et al., 2004; Ebitz and Platt, 2015). The medial frontal cortex (MFC) computes and represents many aspects of performance monitoring (Bonini et al., 2014; Carter et al., 1998; Ebitz and Platt, 2015; Fu et al., 2019; Heilbronner and Hayden, 2016; Ito et al., 2003; Kerns et al., 2004; Pouget et al., 2011; Sajad et al., 2019; Sarafyazd and Jazayeri, 2019; Shenhav et al., 2013; Sheth et al., 2012; Stuphorn et al., 2000; Tang et al., 2016; Ullsperger et al., 2014; Wang et al., 2018a, 2018b), making it a primary substrate for communicating evaluative signals to downstream processes (Miller and Cohen, 2001; Shenhav et al., 2013; Ullsperger, 2017; Ullsperger et al., 2014).

On the one hand, cognitive control involves modulating specific sensory or motor processes involved in the task performed, thus requiring the availability of task-specific information in performance monitoring signals, a form of "credit assignment" (McDougle et al., 2016; Sarafyazd and Jazayeri, 2019). For example, an error made by dialing the wrong number should be distinguishable from an error made by calling an old friend by the wrong name because they require different correction mechanisms. On the other hand, humans excel at performing novel tasks with little prior training – an aspect of flexible behavior that is difficult to study in nonhuman primates, since they cannot be verbally instructed to execute an untrained novel task. This kind of cognitive flexibility requires domain-general mechanisms that abstract from the sensorimotor details of any particular task (Bernardi et al., 2020; Minxha et al., 2020). In a novel setting, errors and conflicts can have unanticipated causes, and generic control mechanisms such as slowing all movement down and increasing arousal are adaptive because they buy time to recruit more resources for domain-specific adaptations to take effect (Ullsperger, 2017; Ullsperger et al., 2014). Downstream processes implementing generic adaptations in arousal, global motor suppression (Danielmeier et al., 2011; King et al., 2010; Wessel and Aron, 2017), and urgency (Cavanagh et al., 2011; Heitz and Schall, 2012; Murphy et al., 2016; Thura and Cisek, 2017) also depend on the availability of domain-general performance monitoring signals to avoid the need to re-learn how to interpret them for every task. Together, these requirements raise the critical question of how performance monitoring signals are represented in the MFC so that they are accessible to inform both domain-specific and domain-general downstream processes. Answering this question requires recording from multiple single neurons in order to characterize the population-level structure of representations in MFC, about which little is known.

Theoretically, a fundamental trade-off exists between representations that support task specialization and generalization (DiCarlo and Cox, 2007; Fusi et al., 2016). Specialization requires that as many different conditions as possible can be differentiated from each other by a downstream process that has access to a large subset of the neurons in the representation ("dichotomies" in the case of pairwise differentiations as implemented, for instance, by a linear classifier). This requirement can be fulfilled by increasing the dimensionality of the representation: if there are as many dimensions as differentiations, all possible dichotomies can be read out, in principle. By contrast, generalization requires low-dimensional representations that abstract away ("disentangle") details specific to a single task (DiCarlo and Cox, 2007; Higgins et al., 2016, 2018). Theoretical work shows that the geometry of neural representations can be configured to accommodate both of these seemingly conflicting needs (Bernardi et al., 2020). Given a representation formatted in this way, linear decoders (which represent conservatively what downstream neurons could read out assuming a feed-forward architecture), depending on how they are trained, can either differentiate between many different conditions (specialization) or generalize across conditions, tasks, and time (abstraction, "cross-condition generalization"). In such a geometry, closely related conditions in each task/context are placed at similar locations on low-dimensional manifolds; these manifolds in turn are approximately parallel to each other to allow generalization. On each manifold, different conditions within a task/context are placed sufficiently apart to allow maximal differentiation (Bernardi et al., 2020). To satisfy these conditions at the population level, the constituent single neurons must be tuned to combinations of several cognitive variables at once ("non-linear mixed selectivity") (Rigotti et al., 2013) . This set of clear theoretical predictions has been supported empirically to some extent by rare empirical studies in monkeys (Bernardi et al., 2020) and humans (Minxha et al., 2020), but has so far not been explored for the important topic of performance monitoring. Here, we examine the hypothesis that neuronal populations in the human MFC represent conflict, error, and estimated control demand in such a format, making them accessible to the many downstream domain-specific and domain-general processes we reviewed above.

Estimation of the statistical likelihood of environmental events is essential for efficient goal-directed behavior (Behrens et al., 2007; Jiang et al., 2015; Shenhav et al., 2013). A key aspect of this process is estimating the probability of encountering a situation where cognitive control will be needed (Jiang et al., 2015; Shenhav et al., 2013). Human participants engage reactive or proactive control depending on whether conflict is likely or not (Braver, 2012; Carter et al., 2000; Logan and Zbrodoff, 1979; Tzelgov et al., 1992). The former strategy is efficient when conflict is rarely encountered, whereas the latter is necessary when conflict occurs often (Braver, 2012). Neuroimaging studies have shown that the MFC encodes contexts that are implicitly defined by conflict probability (Carter et al., 2000), but it remains unknown how knowledge about such implicit contexts is acquired from the 'trial-by-trial' feedback provided by performance monitoring. Motivated by prior results (Behrens et al., 2007; Darlington et al., 2018; Jiang et al., 2015; Shenhav et al., 2013; Sohn et al., 2019), we here examine the hypothesis that human MFC neurons signal and continuously update the probability of encountering a control-demanding situation. Since the type of control triggered by different kinds of conflicts differs, this requires representations that support both domain-specific as well as domain-general readouts. Estimating the expected frequency of each type of trial requires integrating information over the history of trials, offering the unique opportunity to examine the mechanisms whereby representations are maintained and updated over time. We model the trial-by-trial changes of activity of the neurons encoding estimated conflict probability as a Bayesian updating process, in which estimated priors are updated iteratively every time after an action is completed. We show that a novel type of conflict signal appears only after an action is completed and thereby provides the critical information for updating the conflict prior into a posterior.

We recorded single neurons in the MFC while human epilepsy subjects perform two different cognitive control tasks in blocks: the Multi-source Interference task ("MSIT") (Bush and Shin, 2006) and the color-word Stroop task ("Stroop")(Stroop, 1935). The causes of conflict and thus errors are different in the two tasks - stimulus-response spatial incompatibility ("Simon" effect) and/or stimulus conflict ("Flanker" effect) in the MSIT, and stimulus-response incompatibility due to reading colored words in the Stroop task. This allowed us to study how different sources of conflict are encoded within a single task and across tasks, and how errors are encoded across tasks. Subjects were instructed verbally and performed both tasks with little prior practice, which is critical to examine the underlying neural mechanisms (which might be different for extensively practiced tasks as is typically done in animals).

Results

Task and behavior

We recorded well-isolated single units in the dorsal ACC and pre-SMA, which are two areas within MFC associated with different aspects of performance monitoring. Subjects performed two speeded response tasks that require cognitive control: the multi-source interference task (MSIT) and the color-word Stroop task (Stroop) (**Fig. 1a** and **Methods**; Stroop: 593 neurons in dACC and 607 neurons in pre-SMA across 32 participants (10 females); MSIT: 326 neurons in dACC and 412 neurons in pre-SMA in 12 participants (6 females); some patients only performed one of the tasks due to time constraints, **Table S1**). In the MSIT task (**Fig. 1a**, left), conflict arises due to incompatibility between target identity and target location (Simon conflict, or "si" trials) and/or distracting number identity (Flanker conflict, or "fl" trials; trials with both are referred to as "sf" trials). In the Stroop task (**Fig. 1b**, right), conflict arises due to incompatibility between ink color and semantic meaning of the displayed words. In both tasks, sequences of stimuli were randomized, with each type of trial occurring with a fixed probability (In Stroop, 33% of trials had conflict; In MSIT, 15%, 15% and 30% of trials had si, fl, and sf type of conflict, respectively). Subjects were encouraged to respond quickly by an adaptive response threshold (see **Methods**), ensuring maximal task engagement. The different stimulus-response mappings lead to different goal-relevant and irrelevant stimulus features and thus to different kinds of cognitive conflict and reasons for committing errors.

Reaction times (RT) were significantly prolonged in the presence of conflicts, demonstrating the Simon and Flanker effect in the MSIT task (**Fig. 1b**, left; average RT of 0.76s, 0.86s, 0.93s, 1.03s for non-conflict, si, fl, and sf, respectively) and the Stroop effect in the Stroop task (**Fig. 1b**, right; Stroop: 0.76s vs 0.97s for non-conflict and conflict, respectively). We analyzed participants' sequential performance (RT and accuracy) with a Bayesian online learning framework, building on existing models (Behrens et al., 2007; Jiang et al., 2014, 2015). Our models assume that participants iteratively estimate of how likely it is to encounter a certain type of conflict on the next trial. We refer to this variable as the prior for conflict probability (a real number between 0-1 referred to as 'conflict prior'). Since trial sequences were randomized, subjects could not predict with certainty whether the upcoming trial involved conflict or not. However, they could estimate the conflict probability, which is a task parameter set by the experimenter whose value is unknown to the subject a priori. For MSIT, our models estimated two conflict probabilities (one for si, one for fl) at the same time, based on the finding that both conflicts influenced RT (**Fig. 1b**, left). The trial horizon by which past trials ("conflict history") informed the current estimate was dynamically adjusted by a learning rate parameter, which was also estimated online from the data. In order to obtain an individual conflict prior for every subject (even if the trial sequence was identical), we tuned the iterative estimation model by incorporating RT information, using the expectation-maximization procedure described in prior work (Friston, 2002; Jiang et al., 2015). We modelled the RT generation process as a drift-diffusion process (DDM), where the decision variable represents the *difference* in evidence for the target and distractor response; one bound thus represents the correct outcome whereas the other represents the erroneous outcome. This DDM likelihood function for RT is specified with three hyperparameters: decision bound, drift rates, and drift rate bias (Navarro and Fuss, 2009). Conflict prior entered the decision process by biasing the drift rate (Urai et al., 2019) (**Fig. 1c**, right**;** term v_{bias}): the bias term was the scaled Stroop prior for the Stroop model and the sum of independently scaled Simon prior and the Flanker prior for the MSIT model. The same scaling parameters were used regardless of whether the trial had conflict or not because, by definition, the effect of conflict prior started before the trial congruency was revealed.

We estimated the hyperparameters of this model using an expectation-maximization algorithm (see **Methods** for details). Conflict probability was estimated iteratively by updating the current prior with the observed conflict type on each trial using the Bayes' law; the updated conflict posterior then served as the prior for the next trial. This online nature of the model captured how human subjects learned about the statistics of conflict trials as they were experienced sequentially. In the following analyses, we refer to the means of the prior and posterior distributions as conflict "prior" (before stimulus onset) or "posterior" (after action completion; **Fig. 1d** shows an example MSIT session). We considered two alternative classes of models with additional free parameters: 1) models estimating conflict probability (Stroop, Simon or Flanker) using all data at once instead of trial-by-trial updating; 2) reinforcement learning models that perform trial-by-trial updating using a constant learning rate. All of these alternative models required offline fitting using all data. Our RT-tuned Bayesian learning model performed significantly better than either class of alternative models in terms of explaining RT and the conflict sequence (**Tables S2 and S3** for a summary of model comparisons). Additionally, RT tuning significantly improved the Bayesian model in terms of explaining RT (**Table S2**, compare columns "RT tuned" and "no RT tuned"; MSIT delta BIC = -348.5; Stroop delta BIC = -508) and the trial congruency sequence (**Table S3**, compare columns "RT tuned" and "no RT tuned"; MSIT delta BIC = -157; Stroop delta BIC = -232). We thus used the RT-tuned Bayesian model for all neural analyses.

We next examined what aspects of behavior were related to the model-derived regressors (see **Fig. S2** for values of derived hyperparameters). First, in addition to current trial conflict, the estimated conflict prior had a significantly positive (i.e., an increase in RT) main effect on RT in both tasks (**Fig. 1e;** $\chi^2(1) = 6.75$, $p = 0.009$ for Simon and $\chi^2(1) = 6.79$, $p = 0.009$ for Flanker in MSIT; $\chi^2(1) = 28.1$, $p < 0.001$ for Stroop. Likelihood ratio test). The extent to which RT varied with the conflict prior depended on the type of conflict (in the case of MSIT, Simon and Flanker separately), as indicated by a significant negative interaction term ($\chi^2(1) = 12.94$ for Simon and $\chi^2(1) = 14.2$ for Flanker in MSIT; $\chi^2(1) = 33.3$ for Stroop. $p < 0.001$ for all conflict types. Likelihood ratio test). This relation between conflict prior and RT remained significant when trial ID was added as a nuisance variable (**Fig. S1a**), or when the conflict prior was estimated without RT tuning (**Fig. S1b**). These RT effects were replicated by online participants collected using Amazon mTurk as behavioral controls (**Fig. S1d**). Conflict prior was systematically related to errors: when conflict was likely, subjects were *less* likely to commit an error on this trial, suggesting that more control was engaged (**Fig. S1c**; for MSIT we only considered "sf" trials where most errors occurred; significant main effect $\chi^2(1) = 6.81$, $p = 0.009$ for MSIT; significant interaction with non-significant main effect $\chi^2(1) = 18.59$, $p < 0.001$ for Stroop. Likelihood ratio test). Prior work analyzes the influence conflict on the immediately preceding trial has on RT as a signature of cognitive control (Egner and Hirsch, 2005; Kerns et al., 2004). However, the robustness and generality of the conflict adaptation effect varies between studies (Duthoo et al., 2014; Egner, 2007; Schmidt and De Houwer, 2011) and is not the focus of our study. Rather, we here consider conflict learning effects that occur over the span of many trials and that exists independent of conflict adaptation, as shown in prior work (Jiang et al., 2015). In fact, in our data, conflict on the immediately preceding trial provided a poor estimate of conflict probability; compared to our model, previous trial conflict alone explained significantly less variance in RT (MSIT delta BIC = -365.6; Stroop delta BIC = -298.5; **Table S2**, compare columns "RT tuned" and "Prev conflict"), suggesting that our participants incorporated conflict information from multiple trials back. Collectively, these behavioral data from two tasks demonstrate that our models that estimate conflict probability online explained variance in RT and error likelihood, demonstrating a proactive engagement of control.

Neuronal correlates of performance monitoring signals

We focused on three types of epochs for analyses (**Fig. 3a**): baseline before stimulus onset, a 500ms epoch centered at the mid-point between 100ms after stimulus onset and button presses ("ex-ante"), and epochs immediately following button presses ("ex-post"). To assess whether signals relevant for performance monitoring are represented in each epoch, we classified neurons by cognitive variables important for performance monitoring. We identified neurons selective for prior mean or prior variance in the baseline period, for conflict in the ex-ante and ex-post period, and for error, surprise, posterior, and posterior variance in the ex-post period (see single-unit examples in **Fig. 2**; schematic of analysis epochs in **Fig. 3a**; and a summary of overall cell counts in **Fig. 3b**). In MSIT, in order to isolate effects related to the Simon conflict, we refer to the union of "si" and "sf" trials as "Simon trials" and the union of "fl" and no-conflict trials as "non-Simon trials". Similarly, to isolate the effect of Flanker conflict we refer to the union of "fl" and "sf" trials as "Flanker trials", and the union of "si" and no-conflict trials as "non-Flanker trials". Except when noted otherwise, we pooled neurons across dACC and pre-SMA because neuronal responses were similar across areas (**Fig. S3a-b**).

Single units tracked aspects of performance monitoring in both tasks (single-unit examples in **Fig. 2**; summary in **Fig. 3b**). During the baseline epoch, a significant proportion of neurons encoded the mean or the variance of the prior distribution for conflict probability (**Fig. 3b**, blue). In the ex-ante epoch, a significant proportion of neurons encoded conflict (15% in MSIT and 12% in Stroop; Fig. 3b, green), consistent with previous reports (Fu et al., 2019; Sheth et al., 2012). In the ex-post epoch (**Fig. 3b**, yellow), neurons encoded conflict (20% in MSIT; 17% in Stroop), conflict surprise (19% in MSIT; 10% in Stroop), occurrence of errors (22% in MSIT; 19% in Stroop), and the mean and variance of posterior distribution of conflict probability (14/26% in MSIT; 20/12% in Stroop). The signal we refer to as conflict surprise is an unsigned conflict prediction error generated by the experienced conflict given the current prior estimate, a critical component in computing the posterior from the prior (see below). The percentage of units selective for a given variable were similar between the two tasks (**Fig. 3b**).

Many identified neurons showed selectivity for more than one cognitive variable (**Fig. 3c-d**), suggesting a role in bridging different types of information. Approximately 30% of conflict neurons were active *exclusively* in either the ex-ante, early (0-0.5s after button presses) or late (0.5-1.5s after button presses) ex-post epochs (**Fig. 3c**), with some (~12%) active throughout the trial after stimulus onset ("extended"). The distribution of conflict signals across time was strikingly similar between MSIT and Stroop (**Fig. 3c,** compare left and right). We were particularly intrigued by the prominence of neurons signaling conflict ex-post (15-20% of neurons in both tasks; **Fig. 2c** shows an example), which has not been reported before. This conflict signal, whose timing was too late to be useful for within-trial cognitive control, was more prominent compared to the one found in the ex-ante epoch in both tasks (15% vs 20%, $\chi^2(1) = 5.08$, $p = 0.024$ for MSIT; 12% vs 16%, $\chi^2(1) = 9.19$, $p = 0.0024$ for Stroop, chi-squared test). We note that signaling conflict "after the fact" is predicted by our Bayesian conflict learning framework, in which this ex-post conflict serves as the conflict "outcome" signal indicating that the trial was not only *correct* but also with or without conflict, information necessary for computing the conflict posterior from the prior. We found that many conflict neurons also signaled errors, surprise, posterior, or combinations of these variables (for example, signaling conflict, error, and posterior at the same time) (**Fig. 3d**). This multiplexing of signals depended on the timing of conflict signals. The proportion of conflict neurons that also carried information about the posterior (light green bars) increased significantly towards the end of the ex-post epoch, when updating would be most complete and thus the conflict posterior was computed (compare proportion of conflict neurons that multiplexed posterior information in the late ex-post epoch with those that do so in other epochs; $\chi^2(1) = 6.14$, $p = 0.01$ for MSIT; $\chi^2(1) = 6.22$, $p = 0.01$ for Stroop, chi-squared test). Consistent with this idea, the group of neurons signaling conflict exclusively in the ex-ante epoch ("ex-ante conflict only") showed the least multiplexing, indicating a primary role in monitoring conflict during action production (proportion of "pure" conflict neurons active only during the ex-ante epoch vs. those that are active in other epochs; $\chi^2(1) = 5.31$, $p = 0.02$ for MSIT; $\chi^2(1) = 8.78$, $p = 0.003$ for Stroop, chi-squared test). Additional evidence for a differential role of ex-ante and ex-post conflict signals is provided by comparing the point in time when these signals were first available in each brain area. Here, we extracted for each *conflict* trial the point in time when spike train was first significantly modulated for ex-ante and ex-post conflict neurons (using a Poisson spike train statistics-based approach (Hanes et al., 1995)). By this measure, ex-ante conflict information was first available in dACC, followed by pre-SMA (**Fig. 3e**; median difference = 138ms; $p < 0.001$, Wilcoxon rank sum test). By contrast, ex-post conflict information was available first in pre-SMA, followed by dACC (**Fig. 3e**; median difference = 161ms; $p = 0.002$, Wilcoxon rank sum test). This pattern is consistent with a leading role of pre-SMA in post-action performance monitoring (Fu et al., 2019), and a leading role of dACC in conflict monitoring during action production. Collectively, these ex-post neuronal responses appeared to reflect the process of updating internal estimates of conflict probability based on present trial outcome as signaled by conflict and error neurons. We next tested this hypothesis.

Posterior neurons demonstrated the greatest degree of multiplexing (**Fig. 3f**). Only ~18% of posterior neurons signaled posterior exclusively, with the remainder in addition also signaling prior, conflict, surprise, or a mixture of these. This extensive overlap between posterior signals and each of these ex-post constituents might reflect the computation of the conflict posterior, which would involve all these variables. We next tested whether prior neurons (which are selected during the baseline period) changed their spike rates to reflect the updating process in the ex-post epoch (1s after button press). If a neuron correlates with prior on a trial-by-trial basis and the prior is updated into the posterior after each action, the spike rates of this neuron should reflect this updating. As a neural measure of updating, we used the difference of mean-removed firing rates in two epochs: the early ex-post epoch (1s after button presses) and the baseline. As a behavioral measure of updating, we used the numerical difference between posterior and prior means as estimated by the Bayesian models. We then correlated these two trial-by-trial measures for each prior neuron. Across all prior neurons, correlation was significantly positive for all types of conflict priors (**Fig. 3g**, $p < 0.001$, t test against zero. Mean correlations in Simon, Flanker and Stroop are 0.042, 0.032, 0.065, respectively). This result indicates that prior neurons changed their spike rates in the early ex-post epoch, where the conflict outcome was revealed, to reflect the updated posterior. Together, these data demonstrate that a potential role for the ex-post monitoring signals is to update an online estimate of conflict probability.

Event-related potentials that reflect activity of prior cells

The intracranial EEG data recorded simultaneously with the single units revealed an event related potential following button presses on correct trials (**Fig. 3h**; "CRP", or correct-related potentials). Event-related potentials (ERPs) reflect synchronous postsynaptic potentials of cortical pyramidal neurons within the cortical microcircuitry (Buzsáki et al., 2012; Herrera et al., 2020; Woodman, 2010). Similar to the ex-post neurons we investigated, CRPs on average followed button presses, had larger amplitude on conflict trials compared to non-conflict trials (**Fig. S3c**, $\chi^2(1) = 21.05$, $p < 0.001$, likelihood ratio test) and showed an interaction effect between trial congruency and conflict prior (a measure of conflict surprise; $\chi^2(1) = 8.48$, $p < 0.001$, likelihood ratio test), carrying population-level information important for updating conflict prior. We thus hypothesized that these prominent ERPs might represent inputs for the prior neurons recorded simultaneously. We tested whether variance in the spike counts of prior neurons could be explained by the CRP amplitude for each point in time across the trial (mixed-effect Poisson regression models tested with likelihood ratio test, see methods). We investigated dACC and pre-SMA separately, consistent with our previous work (Fu et al., 2019). We found that the activity of prior neurons in both dACC and pre-SMA around button presses was significantly correlated with the CRP amplitude on a trial-by-trial basis (**Fig. 3i-j**, Poisson mixed-effect regression model, which included RT and prior as nuisance variables. $p < 0.01$ for all time bins marked by black dots on top, likelihood ratio test. Multiple comparisons were corrected for using the false-discovery rate method), but with earlier onset in dACC than in pre-SMA (0s vs 0.325s after button presses). This indicates that the CRP amplitude (which occurs in the ex-post period) predicted the activity of prior neurons around button presses on a trial-by-trial basis, revealing a neuronal correlate for this prominent ERP.

Biophysical basis for encoding of prior/posterior

Estimating priors/posteriors in our task necessitates the integration and maintenance of information across multiple trials, a non-trivial property of neural circuitry (Wang, 2002). We therefore investigated whether the functional properties of neurons that encoded priors differed from those that did not. The metrics we used was the temporal correlation profile of baseline spike counts across trials and the width of the extracellular waveform. We chose these properties because the autocorrelation of spike counts at rest of PFC neurons predicts a neuron's participation in working memory (Cavanagh et al., 2018) as well as value coding (Cavanagh et al., 2016), and PFC neurons that encode past reward outcomes have narrower waveforms (Kawai et al., 2019). The timescale of autocorrelation we sought to investigate here is over the span of minutes (multiple trials). To this end, we employed Detrended Fluctuation Analysis (DFA) (Peng et al., 1994) to quantify the self-similarity of baseline spike counts for each neuron, treating the trial-by-trial baseline spike counts as time series data. DFA provides a measure of self-similarity closely related to the slope of the power spectrum (and thus the autocorrelation), but without assuming stationarity. A DFA α value greater than 0.5 indicates a positively correlated process, whereas $\alpha = 0.5$ indicates an uncorrelated process. We found that in both tasks, neurons representing priors had significantly higher DFA α values compared to other categories of neurons (**Fig. 4a-b**, left panels; $p < 0.001$, ANOVA), with DFA α positively correlated with the strength of prior information carried by a particular neuron (**Fig. 4a-b**, right panels; $p < 0.001$, $r = 0.24$ for MSIT, $p < 0.001$, $r = 0.21$ for Stroop, Spearman's rank correlation. Separate data were used for computing these two metrics, see **Methods** for detail).

We next investigated the relation between a neuron's tendency for long-term maintenance of information (as indicated by $\alpha > 0.5$) and its spike width, a biophysical measure that differs between different types of cells (Bean, 2007; Mosher et al., 2020). The relation between DFA α value, autocorrelation, slope of the power spectrum, and spike width can be seen in the two example neurons shown in **Fig. 4c-f**. The orange neuron, which had a $\alpha = 0.91$, had a narrower spike waveform, larger autocorrelation and steeper power spectrum slope than the gray neuron with $\alpha = 0.54$. Across all recorded neurons in both tasks, DFA α values were negatively correlated with spike width (**Fig. 4g-h**; $r = -0.19$ in MSIT, $r = -0.12$ in Stroop, $p < 0.001$ in both cases, Spearman's rank correlation). Neurons encoding conflict prior/posterior, which requires long-term maintenance, in either task had significantly narrower spike waveforms than all other recorded neurons (**Fig. 4g-h**, right; $p < 0.001$, Wilcoxon's rank sum test). Taken together, these data establish that the long-range temporal correlation of baseline spike counts is an intrinsic firing property of neurons that was predictive of the neuron's spike width as well as the encoding strength of conflict prior/posterior. Neurons that represent conflict priors/posteriors appear to be biophysically distinct and of a different cell type from those that do not code such information, due to their systematically different extracellular waveform and firing properties.

Temporal progression of performance monitoring signals

Given the diversity of firing dynamics seen at the single neuron level (**Figs. 2** and **3**), we next examined the temporal dynamics and robustness of performance monitoring signals at the population level using decoding. We used linear classifiers that had access to *all* recorded neurons, which represents a conservative measure of information available to downstream neurons (Fusi et al., 2016). We first focused on within-time decoding (i.e., training and testing a decoder using data collected in the same epoch). Error, conflict, prior and posterior could be decoded reliably on single trials with high cross-validated accuracy (**Fig. 5a-g** and **Fig. S5a-d**, dotted traces or dotted square shows within-time decoding accuracy). In both tasks, the error signal was decodable with high accuracy throughout the whole ex-post epoch, consistent with our previous report of its role in mediating post-error RT adjustments (Fu et al., 2019) (**Fig. 5a-b**, dotted line). Decoding performance of conflict in both tasks peaked first in the ex-ante epoch (**Fig. 5c-e**, dotted line in the green shading) and then again in the early ex-post epoch (**Fig. 5c-e**, dotted line in the orange shading), before gradually decreasing towards the end of the trial. This time course is consistent with a putative role in estimating conflict probability: conflict is first monitored before committing to a response (ex-ante), followed by a representation of the detected conflict as an outcome signal after button press (ex-post).

We next investigated whether the neural code changed over time by using cross-temporal generalization analysis (i.e., training and testing a decoder in different periods of time). We tested the temporal generalization performance of error and conflict decoders trained using data from three defined ROIs: the ex-ante epoch (0.5s; green shading), the early ex-post epoch (0-0.5s after button presses; orange shading) and the late ex-post epoch (0.75-1.25s after button presses; blue shading). In terms of error coding, the early ex-post decoder generalized poorly to later periods (**Fig. 5a-b**; orange line) whereas the late ex-post decoder generalized well across early and late ex-post epochs (**Fig. 5a-b**; blue line). This interesting asymmetry in generalization suggests that there are two groups of error neurons, one signaling errors strongly but transiently, and one signaling error persistently throughout the entire ex-post epochs. In terms of conflict coding, population decoders performed well only within the training epochs but generalized poorly to other epochs for both tasks and all types of conflict (**Fig. 5c-e**), suggesting a dynamic coding patterns for conflict that changed rapidly as the trial unfolds. In particular, the ex-ante decoder did not decode conflict above chance in the ex-post epochs (green traces in **Fig. 5c-e**). This confirms our single neuron findings that the ex-post conflict signals were not simply a continuation of ex-ante conflict signals, but rather signals carried by different groups of neurons at different points in time.

For the population coding patterns of prior/posterior, we took a region of interest (ROI) approach given their slow-varying nature, using the baseline and ex-post epochs for prior and posterior, respectively. Since the conflict priors are continuously valued and differed between sessions, we binned trials using quartiles of conflict prior to aggregate data across sessions (labelling trials by four prior levels). We also binned the trials by quartiles of posterior for posterior-related analyses (labelling trials by four posterior levels). We then trained a linear decoder to differentiate between priors/posteriors of two different quantiles. The prior and posterior decoder could differentiate between all pairs of prior/posterior quantiles with high accuracy (**Fig. 5f-g** and **Fig. S5a-d**; within-ROI decoding, upper or lower triangular matrices enclosed by dotted boxes), with accuracy scaling with the distance between pairs of quartiles (i.e., higher accuracy for differentiating 1^{st} vs. 4^{th} than for 1^{st} vs. 2^{nd} levels). The prior decoders are able to decode all pairs of posterior levels with high accuracy and vice-versa (**Fig. 5f-g** and **Fig. S5a-d**, plots *not* enclosed by dotted boxes), indicating that the representation of prior/posterior is stable across time.

Decoding performance for error, conflict and prior/posterior had similar temporal profiles in both dACC and pre-SMA, but with higher decoding accuracy in pre-SMA (**Fig. S4a-e**). Notably, conflict in the immediately preceding trial could be decoded only weakly (lower than 60% in accuracy) in the baseline, as expected, and in the early ex-post epoch for Stroop conflict (**Fig. S5e-g**). The weak representation of previous conflict is consistent with our observation that the previous conflict alone was a poor predictor of RT compared to the conflict prior (**Table S2**). Together, these data demonstrate that error, conflict and prior/posterior information can be read out trial-by-trial from the MFC population with high accuracy, with dynamic coding patterns for conflict and error and static coding patterns for prior/posterior.

State-space representation of conflict

We have shown robust encoding of each of the four conflict types involved in both tasks (sf, fi, sf, Stroop) separately, leaving open the question of how the different encoding schemes are related to each other. Are the different types of conflict encoded along a common 'conflict' axis or are they encoded separately with no generalization between the types of conflict? We tested this question in the MSIT task, which has three types of conflict (Simon, Flanker, both). We took as a putative common conflict coding dimension the line that, in neural state space, connects the neural state during "sf" and "none" trials (both conflict vs. no conflict; **Fig. 6a-b**, dotted lines). Projecting left-out single trials from all four trial types onto this coding dimension allowed differentiation between all pairs of conflict conditions in the ex-ante epoch (**Fig. 6b**, left). and all pairs but one (si vs. sf) in the ex-post epochs (**Fig. 6b**, right). Importantly, this result holds even when RT was equalized across the four conflict conditions (**Fig. S6c**; si, fl, sf, non-conflict. See **Methods** for RT equalization procedure), suggesting that this conflict coding dimension was independent of trial difficulty for which RT is a proxy (Gratton et al., 1992). We next investigated whether Simon and Flanker conflict encoding is related to each other by projecting the activity of single trials onto the coding dimension formed by connecting, in the neural state space, the mean of Simon (si+sf) with the mean of non-Simon (fl+none) separately for each time bin (and vice-versa for Flanker (fl+sf) vs non-Flanker (si+none)). Data for testing were held out (not used for constructing the coding dimensions). Coding dimensions for one type of conflict allowed decoding of the other type of conflict with high accuracy (**Fig. 6c**; black trace, coding dimension of Flanker tested with Simon vs. non-Simon; gray trace, coding dimension of Simon tested with Flanker vs. non-Flanker). Together, these data demonstrate that within a single cognitive task, the MFC population formed a conflict representation that generalized across two types of conflict while at the same time also allowed maximal separation between the different types of conflict, a geometry that supports both abstraction as well as task-specific specialization (Bernardi et al., 2020).

When two types of conflict coincide on a trial ("sf" trials in MSIT), is the neural state occupied equal to the sum of the two states occupied by the components ("si" and "fl")? In other words, is the Simon and Flanker representation compositional? Perfect compositionality implies that the vectors for the four trial types ("si", "fl", "none", "sf") are coplanar and form a parallelogram, with the "sf" vector being the diagonal and the opposite sides being parallel to each other (**Fig. 6a-b**). We tested this prediction of parallelism (**Fig. 6a-b**, orange edges and blue edges, respectively) using decoding. If the opposing sides are parallel, a decoder trained to differentiate the two classes connected by one edge should be able to decode two classes connected by the opposite edge (and vice-versa). We found that this was largely the case for both ex-ante and ex-post conflict representation: a decoder trained to differentiate "sf" from "fl" trials, which is simply the axis connecting "sf" and "fl" (orange edge in **Fig. 6a**), was able to differentiate "si" from non-conflict trials projected to this axis above chance, and vice versa (**Fig. 6d**, $p < 0.001$ for both the ex-ante and ex-post data, permutation test). The same was true for the other pair of parallel edges (**Fig. 6d**, testing blue edges in **Fig. 6a**; $p < 0.001$ for both the ex-ante and ex-post data, permutation test). The parallelism was not perfect because the decoding accuracy, while above chance, was relatively low (< 70%) compared to the performance reached when decoding individual types of conflict (**Fig. 6c**). This structure of the representation was disrupted on error trials, in which generalization performance dropped significantly in the ex-ante (**Fig. S6d**; for both edges, 68% and 58% on correct trials vs. 56% and 47% on error trials) as well as the ex-post epoch (**Fig. S6d**; for both edges, 55% and 66% on correct trials vs. 51% and 59% on error trials) on error trials. Lastly, we examined which neurons contributed to the deviation that keeps the axes from being perfectly parallel and thus perfectly compositional, which was assessed by the mismatch between the actual location of "sf" and the predicted location by vector addition of fl + si.

Neurons that encoded Simon and Flanker non-linearly (as measured by the F statistic of the interaction term between Simon and Flanker derived from an ANOVA model) contributed the most to the deviation from linear additivity at the population level (**Fig. 6e**, r = 0.74, p < 0.001, for ex-post data; **Fig. S6e**, r = 0.75, p < 0.001, for ex-ante data; Spearman's rank correlation). Collectively these data suggested that in the MSIT task, neural representations of conflict were structured in a compositional way that separated the four conflict conditions in a parallelogram. This geometry was disrupted on error trials, indicating that this representation was behaviorally relevant.

State-space representation of prior/posterior

We reasoned that the conflict prior can be viewed as a state (an initial condition) that is present before stimulus and to which the population returns after completing a trial. To test this idea, we again binned trials using quartiles of prior/posterior of each trial (labelling trials with 4 prior or posterior levels) and aggregated data across the population. Plotting the neural dynamics in a low-dimensional space spanned by three principal components with largest variance explained (PCA is unsupervised and has no access to the ordinal relation between prior/posterior levels) revealed that the variability across different levels of prior/posterior (~8% of variance) was captured mostly by a single axis (PC3s in **Fig. 6f-h**; green dots mark trial start, red dot trial end), which was orthogonal to most of the time-dependent state changes (captured by PC1s and PC2s, ~68% of variance). During the baseline period (green to cyan dots in **Fig. 6f-h**) neural state changed with low speed (**Fig. 6i** and **Fig. S6f-g**), whereas the speed of changed increased significantly after stimulus onset (**Fig. 6i** and **Fig. S6f-g**, red; p < 0.001, paired t-test), eventually returning to baseline near the starting position (red dots in **Fig. 6f-h**). The distance between the four trajectories was kept approximately constant at all time (**Fig. 6i** and **Fig. S6f-g**), consistent with the levels of prior/posterior being states stably maintained at the individual trial level. Remarkably, the state-space trajectories are not only stable but also preserves the ordinal relation between prior/posterior levels: projection values onto the PC that captured the most variance across prior/posterior levels were arranged in an order consistent with the prior/posterior levels, even though PCA did not have access to such ordinal information (**Fig. 6j-l**, see **Legends** for statistics of the multinomial logistic regression). Taken together, the MFC representation of the conflict prior/posterior information is low-dimensional, stable across time and parametric, consistent with the dynamics of line attractors.

Domain-general performance monitoring signals at the population

What was the relationship between the performance monitoring signals we documented separately in the MSIT and Stroop tasks? Did neurons specialize in encoding a given signal in only a given task or did neurons form a domain-general representation across tasks? If the latter, was this representation abstract in the sense that information about the task identity or task-specific conditions (e.g., the different types of conflicts in MSIT) were no longer available? To answer these questions, we next analyzed a subset of data in which we tracked the same neuron in both Stroop and MSIT (see **Table S1** for a tally of recordings). Note that participants had no knowledge of the second experiment they were going to perform when they performed the first (which was either MIST or Stroop), thereby allowing us to ask how two novel tasks with no prior practice engaged MFC. We used demixed PCA (dPCA) (Kobak et al., 2016) on the neural activity recorded across both tasks to identify coding dimensions for error, conflict and conflict prior/posterior which were stable across time and with task information maximally marginalized out. Namely, the goal is to factorize data into performance monitoring signals, non-specific temporal dynamics as well as signals related to task sets (see **Methods** for details). To match the number of conditions between tasks, we picked non-conflict and "sf" trials in MSIT and non-conflict and Stroop trials in the Stroop task to construct the task-invariant conflict dimension. To assess whether the extracted coding dimensions were meaningful statistically, we used them to decode left-out data that were not used to construct these dimensions. To test generalization across tasks, we first projected both left-out training and testing data onto a dPCA dimension, and then classified the testing data in task using training data from the other task.

We found that the dPCA task-invariant coding dimensions identified this way explained between 9-12% of the variance and allowed training of a decoder in one task and testing it in the other with high accuracy (**Fig. 7a-c**; > 80% accuracy for error and conflicts). To quantify the extent of "demixing" of performance monitoring signals from task set information, we computed the angle between the task invariant coding dimensions and their corresponding task dimension. The error dimension supported task-invariant decoding of error throughout the ex-post epoch (**Fig. 7a**; significant clusters see horizonal bars; p < 0.001, cluster-based permutation tests). The angle between the error coding dimension and the task dimension was 94.47° and did not deviate significantly from orthogonality (p = 0.53, tau = -0.032, Kendall rank correlation). The conflict dimension supported task-invariant decoding in the ex-ante and the early ex-post epoch (**Fig. 7b-c**; significant clusters see horizonal bars; p < 0.001, cluster-based permutation tests). The angles between Stroop-Simon conflict coding dimensions and the task dimension were 81.13°, which did not differ significantly from orthogonality (p=0.19, tau = 0.048, Kendall rank correlation). The angles between Stroop-Flanker conflict coding dimensions and the task dimension were 78.6°, which deviated weakly from orthogonality (p=0.02, tau = 0.086, Kendall rank correlation). This task generalizability did not compromise the capacity of this coding dimension to separate different kinds of conflict within MSIT: classifiers could differentiate between 5 out of 6 pairs of conflict conditions with high accuracy (60% - 90%) in both the ex-ante and ex-post periods based on data projected onto the task-invariant dPCA conflict axes (**Fig. 7d**; p values see figure legend, permutation tests). As a control, we repeated all above dPCA decoding analyses using trial conditions equalized by RT (e.g., selecting conflict and non-conflict trials that had similar RTs) and obtained very similar findings (**Fig. S7a-d**). These results suggests that task-invariant representation of error and conflict did not result from the coincidental condition differences in difficulty for which RT is a proxy (Gratton et al., 1992).

Similarly, the representation of conflict priors and posteriors also allowed task-invariant decoding of this information while at the same time differentiation between 94% of pairs of prior/posterior levels in both tasks (**Fig. 7e-f**; p values see figure; permutation tests). None of the coding dimensions for conflict prior or posterior were significantly non-orthogonal with the task dimension (angles and Kendall's tau values see **Legend**; p > 0.2 for all, Kendall rank correlation), suggesting that the coding dimensions identified for conflict prior or posterior were significantly "demixed" from the task set dimension, a complete factorization. Consistent with this, the task dimension support decoding of which task a trial was from with very high accuracy (>90% leave-one-out accuracy, $p < 0.001$ by permutation tests). Together, these data demonstrate that the neural representation of performance monitoring signals in MFC is configured in such a way that it supports generalizability between two different tasks while at the same time also allowing the readout of task-specific information.

Domain-general performance monitoring signals at the single-neuron level

What gave rise to the flexible coding scheme that supported both task-invariant and task-specific readouts as revealed above? Was this a population level phenomenon or did individual neurons encode a given variable reliably in both tasks? To answer this question, we quantified cross-task coding stability for each neuron using linear regression (see **Methods**). To do so, we pooled data from both tasks and regressed firing rates against a performance-monitoring variable (error, conflict, prior or posterior), a task indicator, and an interaction term (performance monitoring x task). The statistical significance for each regressors were determined by an F test. We refer to neurons that had a significant main performance-monitoring effect but non-significant interaction as “task invariant”, and to neurons that had a significant “performance monitoring x task” interaction as “task dependent” neurons. We selected neurons whose response signaled conflict in ex-ante and ex-post epochs, error in the ex-post epoch, prior in the baseline epoch, and posterior in the ex-post epoch.

Out of the selected neurons of each kind (error, conflict, prior), 33-68% were classified as task invariant (**Fig. 7g-j** and **Fig. S7g-k**, red in pie charts). The extent to which a given neuron encoded a performance-monitoring variable by itself (assessed by t statistic for the main effect) and as part of a population (as derived by weight assigned to the neuron by the identified common dPCA coding dimensions) correlated significantly (**Fig. 7g-j**, scatter; see **Legend** for statistics), with the signs of these measures agreeing with each other in most cases (**Fig. 7g-j** and **Fig. S7 g-k**; for cases where signs differed, see **Fig. S7i-j**). Our dPCA analyses marginalized out information about task and time. As a result, neurons selective during either the ex-ante and/or ex-post epochs contributed to the identified common axis and their contribution were thus analyzed separately. On average, neurons identified as “task invariant” or “task dependent” were assigned significantly larger absolute dPCA weights than non-selective neurons ("others" neurons) (**Fig. 7g-j** and **Fig.S7g-k**, dot density plots on the right). While "task-invariant" neurons in many cases had numerically larger dPCA weights than "task-dependent" neurons, on average they did not contribute significantly more to the task-invariant coding dimensions (**Fig. 7g-j** and **Fig.S7 g-k**, dot density plots on the right; see **Fig. S7g-k** for the case where task-invariant neurons contributed significantly more than task-dependent neurons). This illustrates how the diversity of encoding schemes at the single neuron level gave rise to a population-level one-dimensional coding dimension that supports a robust domain-general readout of performance monitoring signals with high accuracy in both tasks (**Fig. 7a-f** and **Fig. S7a-f**). A downstream neuron receiving input from MFC could in theory derive domain-general performance monitoring signals simply by taking a linear sum and thresholding, with the weights equal to those assigned by dPCA.

Discussion

We found that neurons in the human MFC represent estimated control demand (conflict prior) during the baseline, conflict during action production, and conflict outcome and error immediately after an action was performed in both the Stroop and MSIT tasks. A Bayesian conflict learning model that updated conflict probability iteratively after every trial predicted the existence of a novel kind of outcome signal (ex-post conflict signals) and neurons signaling estimated conflict probability, and we identified neurons encoding both of these cognitive variables. Single neurons encoded a diverse array of variables: some neurons encoded conflict, error and conflict prior/posterior in a task-invariant way, some encoded these variables exclusively in one task but not the other, and some multiplexed task information to varying degrees. Neurons with different encoding profiles were randomly and homogenously sampled within MFC with no apparent clustering. Such interdigitated representation patterns would be difficult to detect with fMRI because even one voxel aggregates the activity from hundreds of thousands of neurons. This complexity at the level of single neuron responses precludes a clear interpretation of domain generality or domain specificity and instead requires an analysis of population-level representations in a high-dimensional state-space. The key insight in this study is that the representational geometry of performance monitoring takes advantage of this complex pattern of single neuron responses. Neuronal populations in the human MFC represented performance monitoring signals in a geometry that allows domain-general readout across tasks, while simultaneously also allowing linear decoders to extract task-specific details. Achieving this tradeoff, in turn, requires that the constituent single neurons multiplex information about the different variables required (i.e., show mixed selectivity).

Domain-general performance monitoring

The evidence for domain generality of performance monitoring from neuroimaging studies is mixed. Some studies conclude that representations of conflict are domain-general from spatially overlapping BOLD activation maps, but the extent of such overlap could depend on the statistical thresholding used (Fan et al., 2003; Liu et al., 2004). Using multi-voxel pattern analysis (MVPA) that avoids statistical thresholding, one study finds voxel clusters that simultaneously encode different types of conflict in the superior frontal gyrus (Jiang and Egner, 2014). However, the majority of conflict-encoding voxel clusters, notably those found in dACC, are specific to one particular type of conflict; domain-general and domain-specific clusters are distributed in distinct anatomical locations. Direct comparisons between single-unit recordings and BOLD-fMRI have demonstrated that the former can exhibit multivariate representations that cannot be detected with fMRI (Dubois et al., 2015). Additional difficulty for neuroimaging studies is the variation in human cingulate anatomy, which reduces the overall signal resolution when registering to a common template brain (Crosson et al., 1999; Vogt et al., 1995) (a constraint our work does not suffer from since we mapped anatomy in individual brains). In the present study we find that it is the *same* group of neurons within MFC that form a geometry allowing the readout of both domain-general and domain-specific conflict signals. The conflict representation we discovered not only generalizes between two types of conflict involved within a single task (which was found with MVPA-fMRI as well (Jiang and Egner, 2014)), but also between two tasks with completely different stimuli, response requirements, and task rules. A key component of performance monitoring is the ability to detect action errors without relying on external feedback. This type of self-monitoring is a central component of metacognition (Yeung and Summerfield, 2012). In the case of confidence judgments, which are also metacognitive, fMRI results indicate that the same parts of MFC are involved across different cognitive domains (perception or memory, (Morales et al., 2018)), but it remains unknown whether this also holds for error monitoring. We show that a subset of neurons signal errors in a domain general manner across the two tasks and all types of conflict. At the population level, these domain-general error neurons enabled domain-general readouts of self-monitored error across both tasks (**Fig. 7g**). Future work is needed to demonstrate whether this is also the case for metacognitive signals other than errors.

Domain-specific performance monitoring

We found error neurons independently in both the Stroop task (as previously reported (Fu et al., 2019)) and the MSIT task (a new finding), with a subset of these neurons signaling error only in one or the other task. This would be expected given that what causes an error in the two tasks differs substantially: distraction by the prepotent tendency to read in the case of the Stroop task, and distraction by the tendency to respond to different keys (Simon) or distraction by the flanker number (Flanker) in the case of the MSIT. Importantly, these call for different compensatory mechanisms: for example, Stroop errors can be compensated for by suppressing attention to the word meaning, Simon errors by suppressing attention to the spatial location of the target, and Flanker errors by suppressing attention to the flanking distractors. The same argument applies to the conflict signal; together, these considerations highlight the fact that cognitive control requires information about performance specific to the task performed. Consistent with this requirement, the MFC neural states varied both along a generic conflict dimension that generalized across tasks as well as along all six types of specialized conflict dimensions. By changing connection weights, downstream neuronal processes can flexibly access performance monitoring signals at both extreme as well as intermediate levels of abstraction and drive behavioral adaptations accordingly. Prior work demonstrate the MFC's causal role in credit assignment within a single task by showing that macaque MFC neurons are involved in attributing the cause of an error to either low-level perceptual noise or exogenous changes in the response rule (high-level) (Sarafyazd and Jazayeri, 2019). We found that both at the level of single neuron and population activity, error and conflict signaling offered a robust readout of the task in which these performance disturbances were experienced, even when the subject had no prior task exposure. This task specificity at the population level is supported by the fact that some MFC neurons which did not signal errors or conflict in a previous task started to do so in a novel task with no prior training (**Fig. 7g-I** and **Fig. S7g-h**). These results are broadly consistent with the MFC's role in credit assignment within a task and demonstrate the remarkable flexibility of the MFC performance monitoring circuitry.

Compositionality of conflict representation

Further insight into how representations can be both general and specific is offered by examining the activity during "sf" conflict trials (both Simon and Flanker conflict) in the MSIT task. The potential compositionality of the representation of two kinds of conflict can be formulated as a generalization problem: if Simon and Flanker conflict are linearly additive, decoders trained to identify the presence of only Simon or Flanker conflict should generalize to the "novel" situation where both types of conflict are present. We found that this was the case, with the neural state approximately equal to the linear vector sum of the two neural states when the two types of conflict are present individually. This suggests that conflict representations are additive to a large extent (with the extent of deviation predicted by the degree of nonlinear mixing present). The (approximate) factorization of conflict representation is important for both domain-specific and domain-general adaptation: when different types of conflict occur simultaneously and the representation can be factorized, downstream processes responsible for resolving each type of conflict can all be initiated. On the other hand, domain-general processes can also read out the representation as a sum and initiate domain-general adaptations.

Estimating control demand enabled by ex-post conflict neurons

We found that subjects' performance in two conflict tasks was best explained by models that iteratively estimate how likely the next trial is to contain each possible type of conflict. These models integrate information across many trials and outperform the conflict adaptation model whose horizon only includes one trial back in terms of estimating and predicting control demand. We modelled the abstract decision process (which predicts RT and correct/incorrect, but not the actual choices) as a drift diffusion process with different drift rates on a conflict and non-conflict trial. Conflict prior was incorporated by adding a bias to the drift rates. The choice of DDM is informed by prior work demonstrating that sequential effects in perceptual decisions are modelled best with drift rate biases (rather than biases in other factors such as starting point or boundary) (Urai et al., 2019). We chose a Bayesian DDM framework because: 1) it estimates control demand (conflict probability) iteratively as our participants did and is thus neurally feasible; 2) similar models have found success in explaining behavior in cognitive control tasks

(Ide et al., 2013; Jiang et al., 2015); 3) it provides trial-by-trial regressors for neural analyses. Notably, this model performed significantly better than one assuming that subjects try to estimate a fixed conflict probability (which is how the task is designed). This shows that subjects were sensitive to the random variability in the trial congruency sequence to adjust their response strategy as reflected in their response times.

A prediction of our Bayesian learning framework is that during the ex-post period, neurons would signal whether the just completed trial was a conflict or not. This is because this kind of "after the fact" conflict signal is needed to compute the posterior from the prior. Confirming this prediction, we found two kinds of conflict signals: one that occurred after action completion as predicted ("ex-post"), and one that occurred during action production ("ex-ante") as expected (Botvinick et al., 2001; Sheth et al., 2012). Separate groups of neurons gave rise to these two types of conflict signals. To the best of our knowledge, this ex-post coding of conflict is a novel kind of conflict signal not previously documented. We posit that the ex-post conflict signal is an outcome signal (Shenhav et al., 2013) that is used for updating slowly varying representations of estimated conflict probability. Interestingly, there is significant overlap between error neurons and these ex-post conflict neurons. Confirming this, we found that a common coding axis exists that supports decoding of both error and conflict, though the decoding accuracy is significantly lower for conflict than for error (**Fig. S6a-b**). These results suggest the origins of ex-post conflict and error signals may be similar: a putative prediction error computed based on an efference copy (Lo and Wang, 2006). Future work is needed to test this new hypothesis. We also revealed a direct neural correlate of the updating step: firing rate changes of prior neurons during the ex-post period are positively correlated with prior-posterior differences in the model. Altogether, the neuronal responses we found fit remarkably well to the parameters of a Bayesian model that used trial-wise updating to estimate upcoming conflict – a critical ingredient in the control of flexible behavior in changing environments.

Prior neurons as a substrate for proactive control

The dual mechanisms of control (DMC) framework (Braver, 2012) distinguishes between proactive and reactive control and suggests that proactive control involves sustained anticipatory activity. We posit that the prior-signaling neurons reported in this study can inform proactive control processes about the estimated demand for control in the upcoming trial. Compatible with this view, responses of these prior/posterior neurons in the MFC were sustained and stable across time. At the population level, the prior/posterior representations were parametric with dynamics indicative of a line attractor, from which neural activity departs and then returns to when completing a trial. The distance in neural state space between neural trajectories remained stable across the trial, as expected from a signal that only varies slowly and that reflects a learning process that occurs over multiple trials.

A likely contribution to the temporal stability of prior coding are the biophysical properties of prior neurons, which differed from other neurons in two ways. First, these neurons exhibited long-range autocorrelations in their baseline spike counts. While spike-count autocorrelations in the range of seconds are known to differ between brain regions (Bernacchia et al., 2011) and between neurons with different tuning (Cavanagh et al., 2016, 2018), here we examined long-range temporal correlation on a substantially longer timescale (minutes) by assessing the self-similarity of trial-by-trial spike counts during the baseline using Detrended Fluctuation Analysis (DFA) (Hardstone et al., 2012). Our finding that DFA values for prior-encoding neurons are high (indicating long-range temporal autocorrelations) suggests that they are ideal substrates for representing a slowly varying internal state. In addition, these neurons tended to have shorter extracellular waveforms. While this relationship is complex (Vigneswaran et al., 2011), neurons with narrower spikes are more likely to be interneurons and the functional role of thin- and broad- spike MFC neurons is often different (Bean, 2007; Sajad et al., 2019). Notably, in macaques, thinner-spike neurons are more likely to have long autocorrelations making them ideal to carry slowly changing information across trials (Kawai et al., 2019). Our findings tie together the biophysical properties of single neurons with their tuning, indicating that conflict prior neurons are ideally suited to carry slowly changing information across trials.

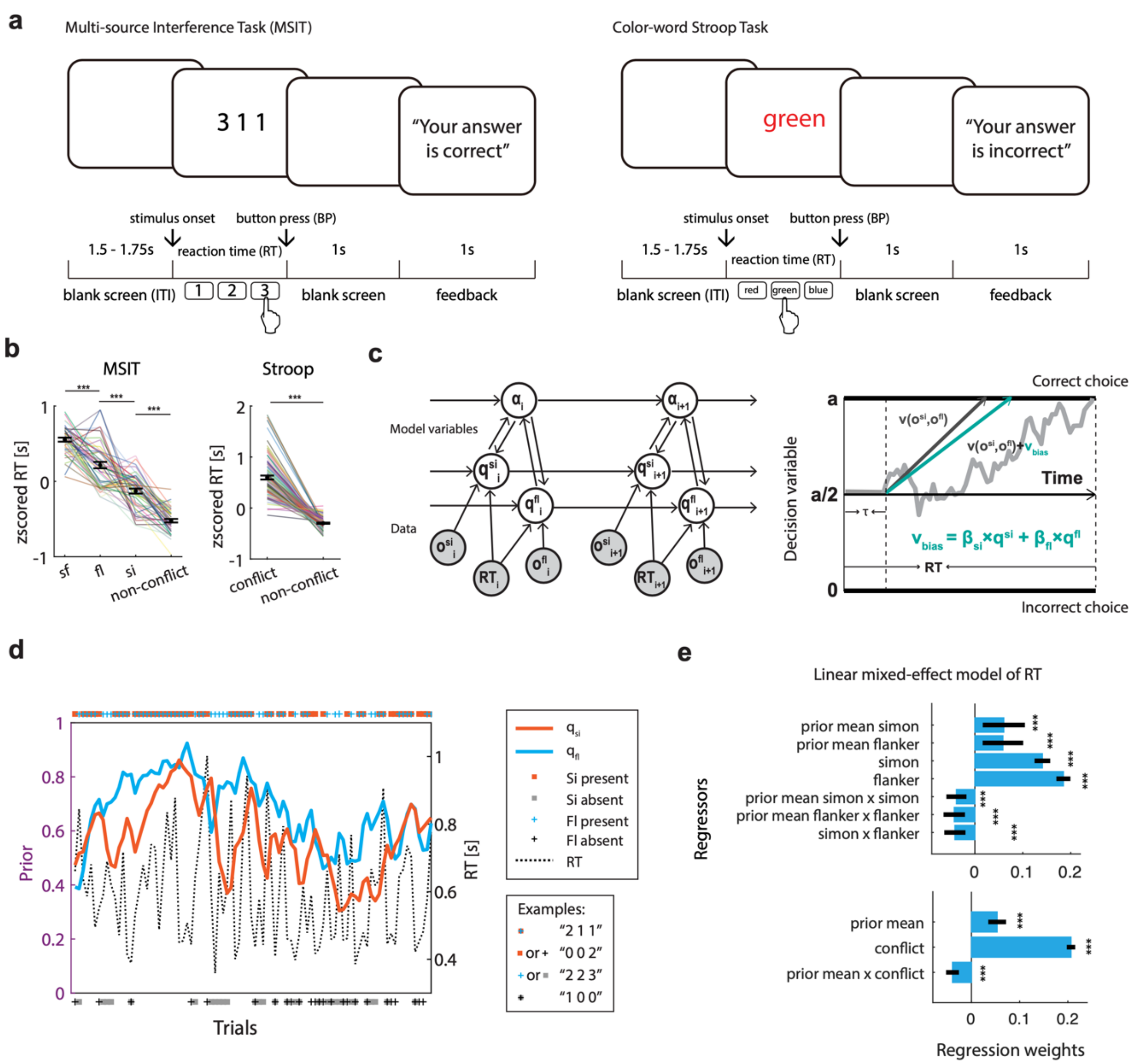
Acknowledgments: We thank the members of the Adolphs and Rutishauser labs and Lawrence J. Jin for discussion. We thank all subjects and their families for their participation and the staff of the Cedars-Sinai Epilepsy Monitoring Unit for their support.

Funding: This work was supported by NIMH (R01MH110831 to U.R.), the NIMH Conte Center (P50MH094258 to R.A. and U.R.), the National Science Foundation (CAREER Award BCS-1554105), the BRAIN initiative through the NIH Office of the Director (U01NS117839), and the Simons Foundation Collaboration on the Global Brain (RA, ZF).

Figures

Figure 1. Tasks, model, and behavioral results.

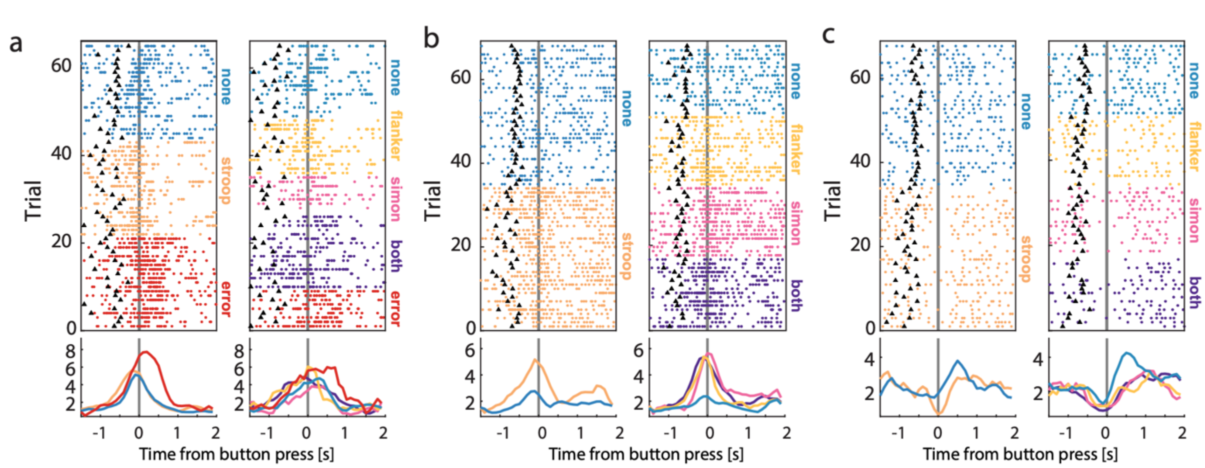
Figure 1



(a) Tasks. MSIT (left) and Stroop (right). Participants indicated the identity (1,2, or 3) of the unique number in MSIT, and the ink color of a colored word in Stroop. Feedback followed 1s after responses and indicated trial outcome.
(b) RTs were significantly prolonged by conflict in MSIT (left) and Stroop (right), showing the Simon, Flanker and Stroop effects. Each line is a session (N = 41 and 82, respectively).
(c) Graphical representation of the updating process (left) and the decision process modelled as a drift diffusion process (right). Incorporating RT likelihood function (DDM) allows the tuning of model estimate for conflict probability. Shown is the MSIT model, which has the five variables volatility (α), predicted Simon conflict (q_{si}), predicted Flanker conflict (q_{fl}), observed Simon conflict (o_{si}), observed Flanker conflict (o_{fl}), and RT. Observables (conflict type, RT, and outcome) are shown in gray, internal variables in white. Arrows indicate information flow. As the trial started, the volatility variable is updated first, and then both predicted conflicts are updated by the respective observed conflicts (Bernoulli likelihood) and RT (DDM likelihood). A linear combination of the Simon conflict and Flanker conflict priors on each trial are entered as a drift rate bias. The hyperparameters for RT tuning included the two linear coefficients before conflict priors (β_{si}, β_{fl}), boundary separation (a) and the four separate drift rates for the four conflict conditions (si only, fl only, si+fl, no-conflict). The conflict priors and posteriors were used as regressors for subsequent behavioral and neural analyses.
(d) Estimated mean of the prior for Simon probability (orange) and Flanker probability (blue) from an example MSIT session. Markers placed on the top indicated that either Simon conflict (orange square) or Flanker conflict (blue cross), or both, was *present* on a trial. As is shown here, the priors increase when there is a run of conflict (left part of the graph, both blue and orange traces go up).
(e) Regression analyses of RT using linear mixed-effect models. Blue bars show regression coefficients; black bars show confidence interval. All regressors explained significant variance as determined by the likelihood ratio test (see **Methods**). Conflict prior positively predicted RT in MSIT and Stroop.
*p < 0.05, ** p < 0.01, *** p < 0.001, n.s., not significant (p > 0.05).

Figure 2. Example neurons in Stroop (left) and MSIT (right).

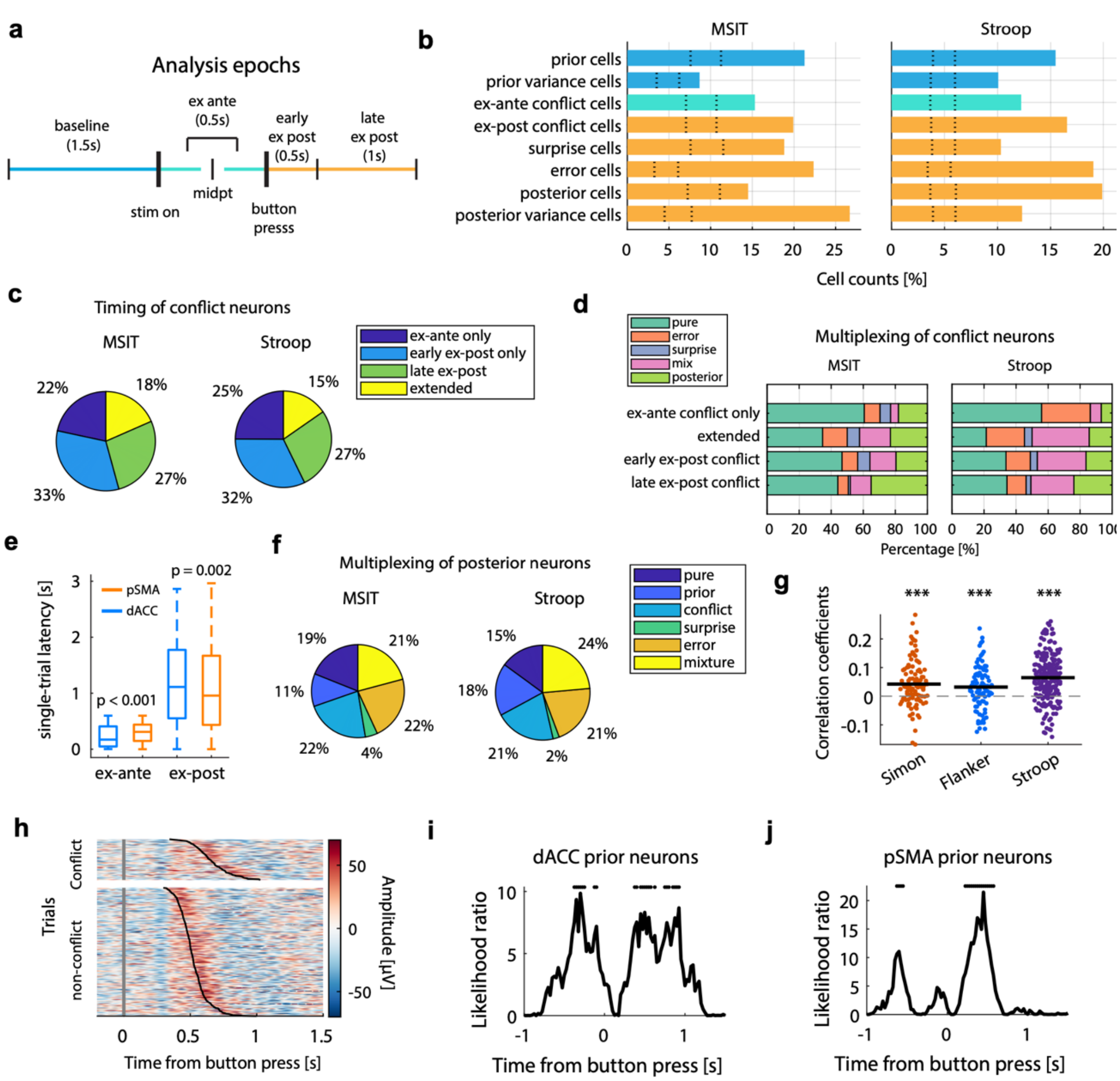
Figure 2



Shown here are raster plots and peri-stimulus time histograms for three example neurons in both tasks. Left panel shows data from Stroop, right show data from MSIT. Data are aligned to button presses (t=0). The black triangles mark stimulus onset.
(a) Neuron signaling action errors.
(b) Neuron signaling conflict by firing rate increase around button presses.
(c) Neuron signaling conflict by a firing rate decrease around button presses.
Trial types are marked by colored words on the right side of the box. These example neurons show similar responses dynamics in both tasks. Trials were re-sorted into groups for display purposes only.

Figure 3. Neuronal selection and ERP analysis.

Figure 3

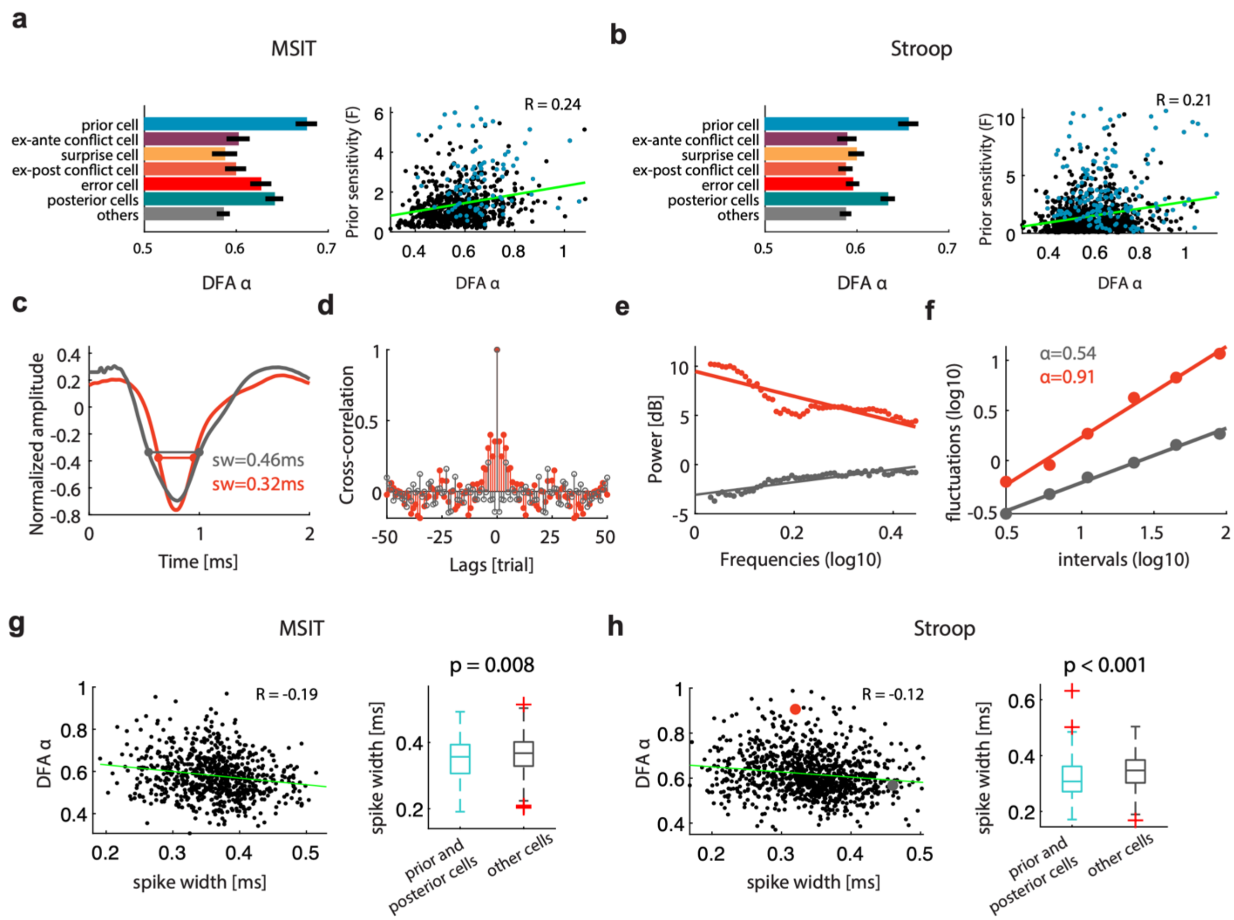


(a) Analyses epochs used in neural analyses. The ex-ante epoch is defined as 0.5s epoch centered at the midpoint between 100ms after stimulus onset and button presses. This was the moment when conflict should achieve its maximum. Ex-post epochs are defined as epochs following button presses. Epochs are colored by blue (baseline), cyan (ex-ante) and orange (ex-post). The thickened vertical bars represent actual physical events, the slim vertical bars demarcate epochs. Note that we also sometimes use 1s after button press as an analysis window.
(b) Percentage of neurons that encode task variables in MSIT (left) and Stroop (right). The color code represents the epochs used to select these neurons (see (a)). Rows are arranged from top to bottom in temporal order. Dotted lines represent 2.5^{th} and 97.5^{th} percentiles of the null distribution obtained from permutation. For all groups shown, $p < 0.001$.
(c) Percentage of conflict neurons that are active exclusively in four groups: ex-ante, early ex-post, late ex-post, and throughout ex-ante and ex-post epochs.
(d) Percentage of conflict neurons that were also selective for error, surprise, posterior, or any combination of these factors ("mix"). Substantial proportions of conflict neurons multiplex error and posterior information. The intersection between conflict and posterior increases towards the later part of the trial. Rows are arranged from top to bottom in temporal order.
(e) Comparison of single-trial neuronal response latency of conflict neurons in dACC and pre-SMA. Ex-ante conflict neurons become active earlier in dACC than in pre-SMA, whereas ex-post conflict neurons become active earlier in pre-SMA than in dACC. Only correct conflict trials are used in this analysis.
(f) Percentage of posterior neurons that intersect with prior, conflict, surprise and error signaling. We hypothesize that the extensive overlap between these groups reflects posterior computation.
(g) Neuronal signature of updating conflict prior based on the posterior. Correlation is computed between the difference between prior and posterior (behavioral update) and the difference between demeaned $FR_{ex\text{-}post}$ and $FR_{baseline}$ (neural update) for all prior neurons. On average, the correlation is significantly positive, suggesting that the change in firing rates is commensurate with the extent of updating derived from the behavioral model.
(h) An example session of intracranial EEG in Stroop, aligned to stimulus onset (grey vertical bars) and sorted by RT (black lines). Color code represents amplitude in micro volt. An event-related potential, named correct-trial potential (CRP), occurs shortly after button presses that is present on both conflict and non-conflict trials.
(i) Relation between CRP amplitude and spiking activity of prior neurons. Both data were simultaneously recorded in dACC. Likelihood ratio computed by comparing the full Poisson regression model with CRP as a fixed effect with a reduced model without the CRP term, and is plotted as a function of time. Black dots on top mark significant time bins, corrected for multiple comparisons using the false discovery rate (FDR) method.
(j) Same as in (i), but for pre-SMA data.

Figure 4. Long-range temporal correlation autocorrelation of spiking of prior neurons.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.08.451594; this version posted July 8, 2021. The copyright holder for this preprint (which

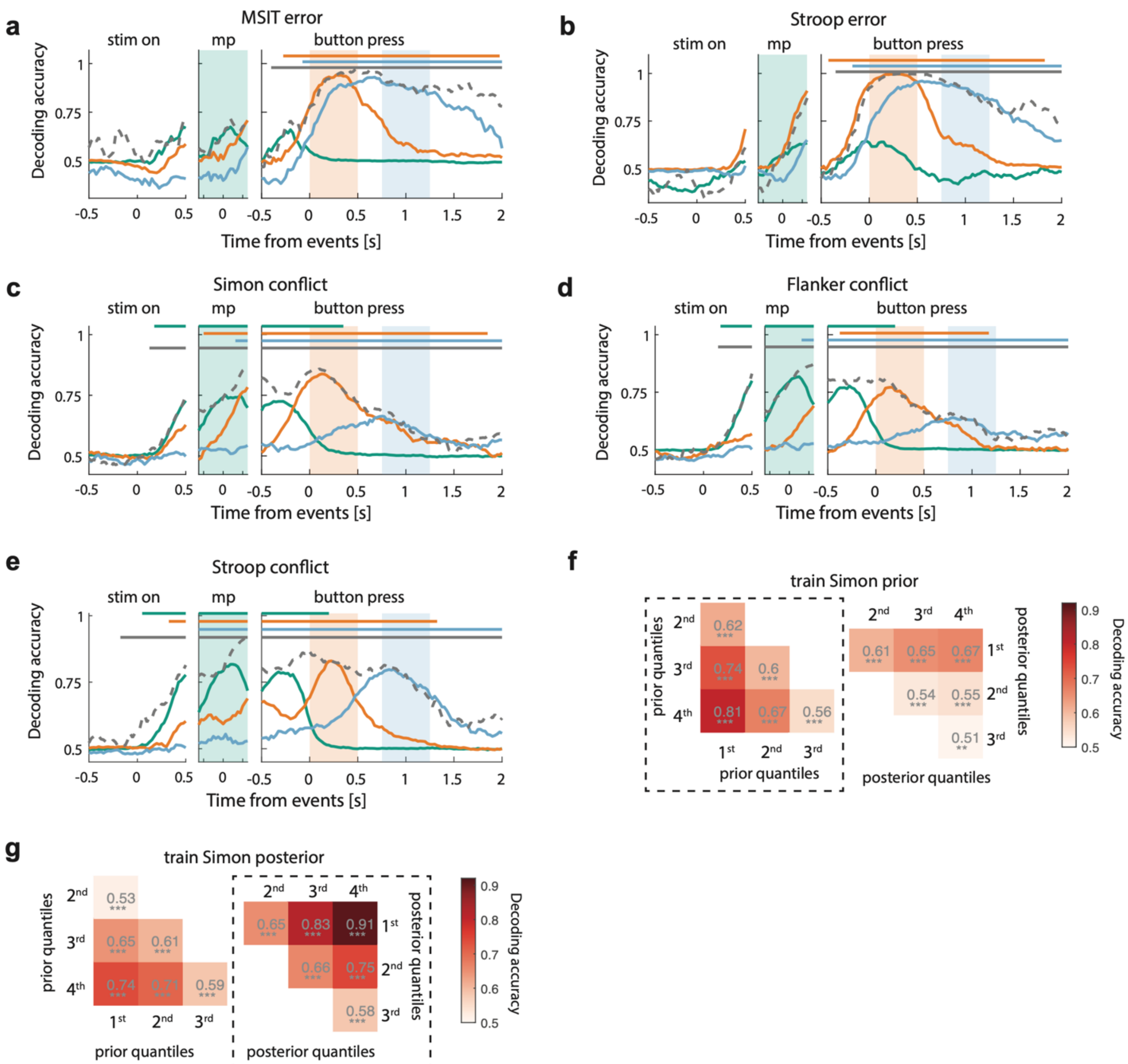
Figure 4



(a) Mean baseline DFA α value for different groups of neurons (left), and correlation between baseline DFA α value and the coding strength of prior, which is assessed by the F statistic computed from regressing prior against baseline spike counts (right) for MSIT. Separate data were used to compute α value and prior coding strength to avoid selection bias. Prior neurons have significant larger baseline DFA α value than any other groups ($p < 0.001$, ANOVA). The coding strength of prior was correlated strongly with the baseline α value ($r = 0.24$, $p < 0.001$).
(b) same as in (a) but computed for the Stroop task.
(c-f) Two example neurons, showing (c) waveforms, (d) autocorrelation, (e), power spectrum, and (f) fluctuations as a function of time intervals used to compute DFA α value (slope). The neuron with narrower spike width has higher DFA α value ($r = -0.19$, $p < 0.001$).
(g) DFA α value is negatively correlated with spike width for MFC neurons (left) in MSIT. Prior and posterior neurons as a group have significantly narrower spikes than other neurons (right).
(h) same as in (g) but for the Stroop task.

Figure 5. Temporal dynamics and cross-time generalization.

Figure 5

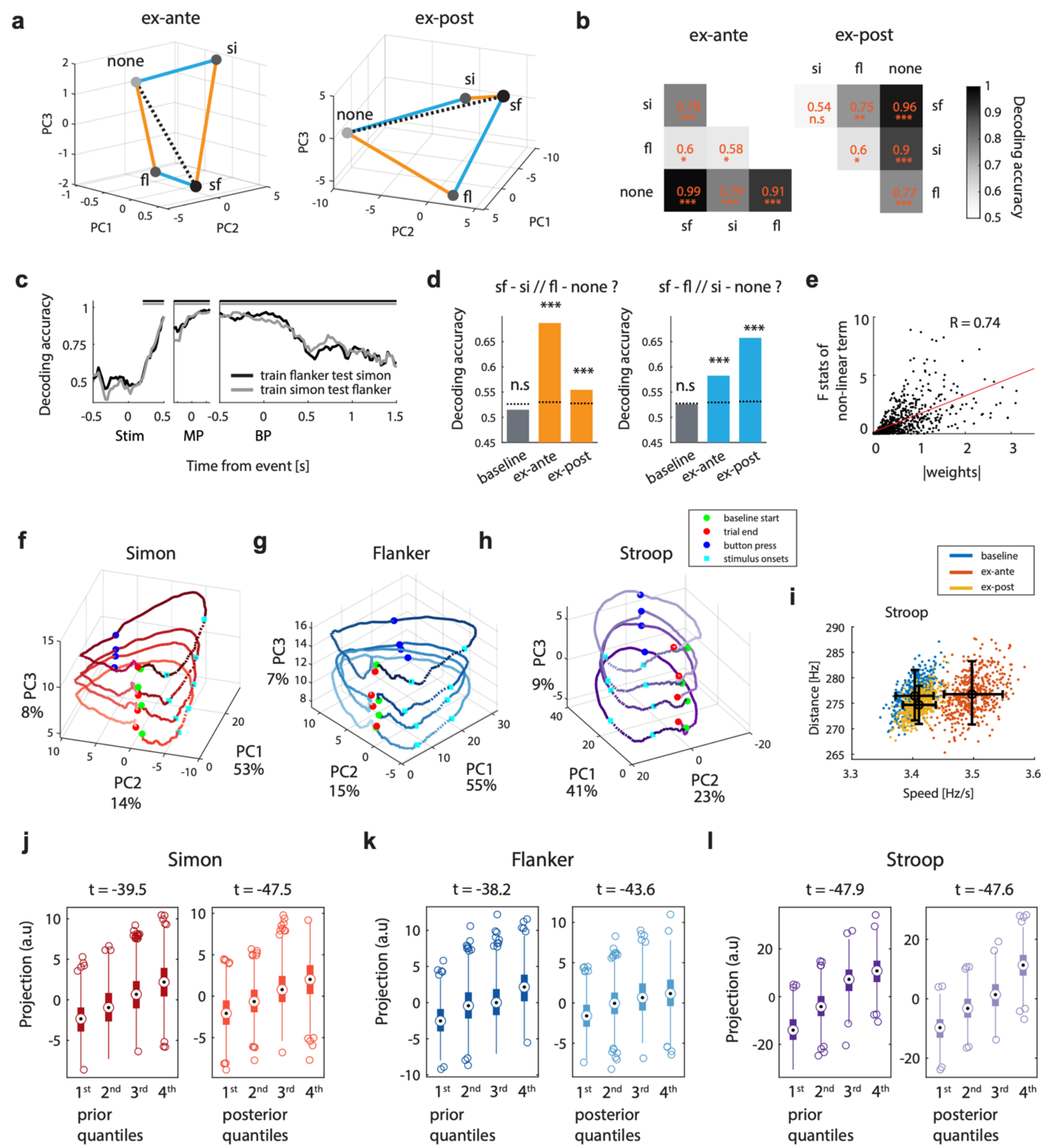


(a-e) Decoding accuracy as a function of time for Stroop error, MSIT error, Simon conflict, Flanker conflict and Stroop conflict. The three panels show data aligned to stimulus onset, midpoint between 100ms after stimulus onset and button presses, button press onset. Dotted gray trace represents within-time decoding accuracy, i.e., the data from the same epoch were used to train and test a decoder. Green, blue and orange traces represent decoding accuracy of decoders trained with data from the epochs demarcated by shading with the same colors, which is a test of the temporal generalization of these decoders. Horizontal bars demarcate the extent of significant clusters in time as determined by the cluster-based permutation test ($p < 0.5$).
(f-g) Decoding accuracy for classifying between pairs of Simon prior or posterior levels (binned by quartiles). Color bars show decoding accuracy. Dotted frames mark the within-time decoding results. Decoders that classify prior quartiles are trained using baseline spike counts (1.5s before

stimulus onset), whereas decoders that classify posterior quartiles are trained using ex-post spike counts (2s after button press). To test temporal generalization of these decoders, prior-trained decoders are tested with posterior data and labels, and vice versa. Dashed boxes represent within-time decoding.

Figure 6. State-space representation of conflicts, prior, and posterior.

Figure 6



(a) Visualization of the conflict population representation in MSIT. Trial mean of four MSIT conflict conditions, Simon only ("si"), Flanker only ("fl"), Simon and Flanker both present ("sf"), and non-conflict ("none"), plotted in space spanned by three principal components. Left panel uses ex-ante data. Right panel uses ex-post data. The extent of compositionality of conflict representation is tested by condition generalization of decoding in (d). Dotted line is the vector used to classify pairs of conflict conditions in (b).

(b) Decoding accuracy of pairwise classification of conflict conditions. Training data and left-out testing data from the four conflict conditions are projected to the population vector flanked by averages of non-conflict trials and sf trials, shown as the dotted line in (a). Color code represents decoding accuracy. This coding dimension separates the four conflict conditions well.

(c) Abstract conflict signal. The three panels show decoding accuracy using data aligned to stimulus onset, midpoint between 100ms after stimulus onset and button press onset, and button press onset. For each time point, a decoder is trained with Simon (union of si and sf) vs. non-Simon (union of fl and none) trials or Flanker (union of fl and sf) vs non-Flanker (union of si and none) trials and used to classify left-out data of Flanker vs non-Flanker trials (grey) or Simon vs. non-Simon trials (black).

(d) Testing compositionality of conflict representation with condition generalization of decoding. If compositional, the sum of the representations of fl and si (vectors none -> fl, none -> si) should be equal to the representation of sf (none -> sf), and the four condition means should form a parallelogram. We tested the condition generalization using raw spike count data from the ex-ante and ex-post (1s after button presses) epochs and from the baseline as a control without dimensionality reduction. Data from the means connected by one of the blue edges in (a) were used to train a decoder, which was then tested with left-out data from the means connected by the opposite blue edge, and vice versa. Blue bars show decoding accuracy (baseline 0.51, ex-ante 0.69, ex-post 0.55). Same was also tested with data connected by the orange edges. Orange bars show decoding accuracy (baseline 0.53, ex-ante 0.58, ex-post 0.66). Both the blue and orange pairs of opposing edges supported such generalization simultaneously as indicated by the above-chance decoding accuracy ($p < 0.001$), demonstrating parallelism and thus the compositionality of conflict representation. Dotted lines show 97.5^{th} percentile of the null distribution from permutation.

(e) Single neuron with nonlinear coding of Simon and Flanker conflict contribute to deviation of conflict representation from perfect linearity. Data used here are from the ex-post epoch. Nonlinear coding of conflict by a single neuron is measured by the F statistic of the interaction term between Simon and Flanker conflict in an ANOVA model with spike counts as the dependent variable. Each neuron's contribution to the deviation from linear additivity in the high dimensional neural space is quantified by the weight of the difference vector between "sf" and "si + fl". Scatter plot shows the relation between these two measures. Red line shows the linear fit.

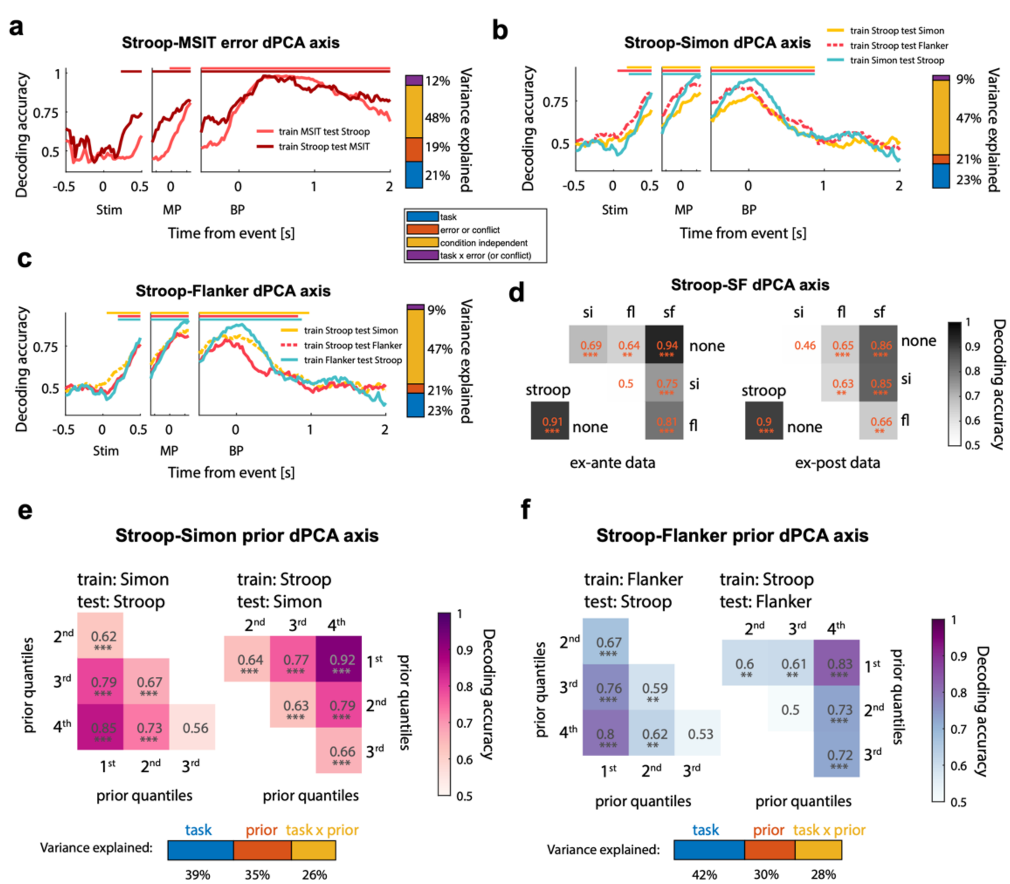
(f-h) Visualization of prior/posterior population representation in MSIT and Stroop. Green dots mark the onset of trial baseline, cyan squares mark the range of possible stimulus onsets, blue dots mark button press and red dots mark end of trial. The range of stimulus onsets (a range because trials are aligned to button press onsets) is shown as broken lines for each prior level. For the portion before button presses (blue dots), the four trajectories correspond to the mean of trials grouped by quartiles of prior. For the portion after button presses, the four trajectories correspond to the mean of trials grouped by quartiles of posterior for after button presses. Trials are aligned to button press onset. The color of trajectories fades as the trial progresses towards the end (dark and light colors correspond to start and end of the trial). Most of the variance related to prior/posterior is captured by PC3, which is orthogonal to most of the time-dependent dynamics.

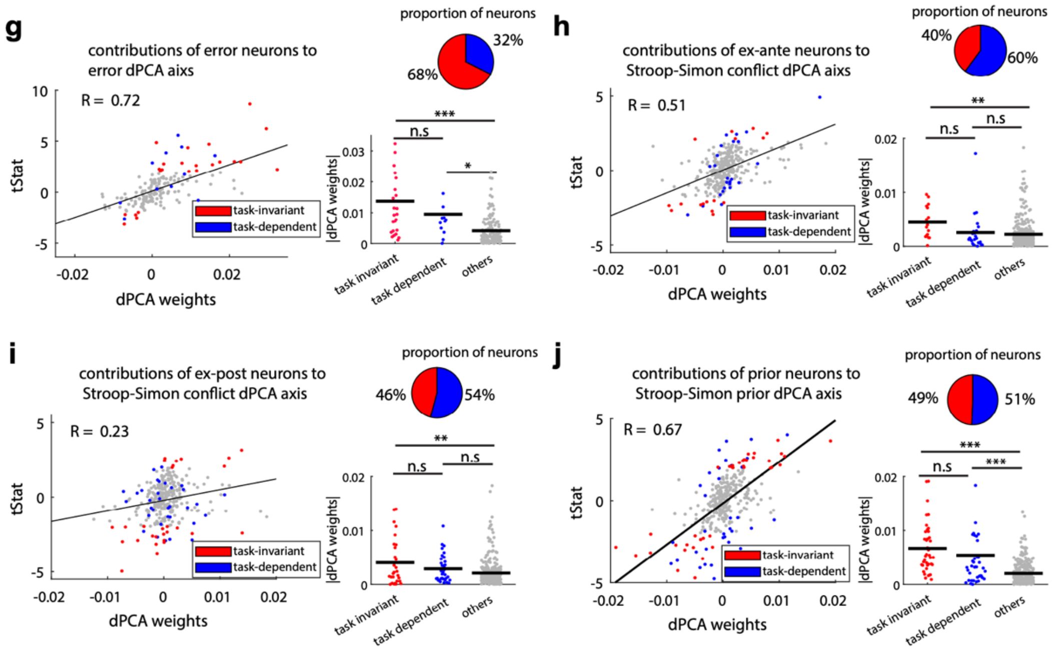
(i) Distance between trajectories and average speed computed from trials grouped by quartiles of Simon conflict prior in the baseline (blue) and the ex-ante (orange) epoch, and trials grouped by Simon conflict posterior in the ex-post epoch (yellow). Trajectories are visualized in Figure 6f. The state space speed stays low during baseline, increases significantly during the ex-ante epoch and decreases back to a value similar to that during the baseline. Distance between trajectories is stable across time.

(j-l) Projection values on the coding dimension of Simon (J, $p < 0.001$ for both prior/posterior, prior $t(11996) = -39.5$, posterior $t(11996)=-47.5$), Flanker (K, $p < 0.001$ for both prior/posterior, prior $t(11996) = -38.2$, posterior $t(11996)=-43.6$) or Stroop (L, $p < 0.001$ for both prior/posterior, prior $t(11996) = -47.9$, posterior $t(11996)=-47.6$) prior or posterior (PC3 in f-h). The order of projection values is on average consistent with the order of the prior/posterior quartiles, even though PCA does not have access to the order information.

Figure 7. Domain-general representation of performance monitoring signals.

Figure 7





(a) Task-invariant decoding of errors in both MSIT (salmon) and Stroop (crimson). The task-invariant coding dimension is extracted using dPCA that marginalizes out task information and time. This dPCA coding dimension is extracted from the error contrast in Stroop (error conflict vs. correct conflict trials) and the error contrast in MSIT (error "sf" trials and correct "sf" trials). This controls for trial conflict and isolates effects related only to error. Left, Accuracy for decoding errors as a function of time. Bar on the right shows the variance explained by the different dPCA components (color code see figure legend). The angle between the error dPC and the task dPC drived from dPCA is 94.47° and did not differ significantly from orthogonality ($p = 0.53$, tau = -0.032, Kendall rank correlation).

(b-c) Task-invariant decoding of conflict in both MSIT (Simon, yellow; Flanker red) and Stroop (green). Because MSIT has two conflict conditions, Simon and Flanker, task invariance was investigated between Stroop/Simon and between Stroop/Flanker conflicts separately. This dPCA coding dimension is extracted from conflict and non-conflict trials in Stroop and either from Simon and non-Simon trials (b) or from Flanker and non-Flanker trials (c), by marginalizing out task and time. The angle between the Stroop-Simon conflict dPC and the task dPC is 81.13° and did not differ significantly from orthogonality ($p = 0.19$, tau = 0.048, Kendall rank correlation). The angle between the Stroop-Flanker conflict dPC and the task dPC is 81.13° and is significantly but weakly non-orthogonal ($p = 0.018$, tau = 0.087, Kendall rank correlation). Left-out conflict trials and non-conflict trials in Stroop, and left-out Simon, non-Simon, Flanker, non-Flanker trials in MSIT are projected and classified by this coding dimension. Left, decoding accuracy of conflict as a function of time. The bar on the right represents variance explained by the different dPCA components (color code see figure legend).

(d) Testing separability of conflict conditions in Stroop and MSIT using data from the ex-ante (left) and ex-post epochs (right). The dPCA coding dimension used in this analysis is extracted by using conflict and non-conflict trials in Stroop and sf and non-conflict trials in MSIT by marginalizing out task information. Because data from ROIs are used, the time dimension is already marginalized out before entering dPCA algorithm. This coding dimension support classification of 83% of pairs of MSIT conflict conditions (upper triangle matrices) as well as Stroop conflict (left corner). Color coding represents decoding accuracy, orange numbers indicate the numerical values of decoding accuracy of that pair of conflict conditions (e.g., the accuracy is 0.69 for decoding si vs. none). Conflict monitoring is thus task-invariant but still preserves maximal separability of task-specific conflict conditions (MSIT).

(e) Task-invariant decoding of all pairs of conflict prior levels in Stroop (lower triangle matrix) and Simon (upper triangle matrix). The dPCA coding dimension here is extracted by using the Stroop conflict prior contrast (the 1st vs. 4th quartiles of Stroop conflict prior) and the Simon conflict prior contrast (the 1st vs. 4th quartiles of Simon conflict prior), marginalizing out task information. Color code represents decoding accuracy. Bar at the bottom shows variance explained of dPCA components (for decoding, the component labelled as "prior" is used). The angle between Stroop-Simon conflict prior dPC and the task dPC is 106.42° and did not differ significantly from orthogonality ($p = 0.78$, tau = -0.01, Kendall rank correlation). The angle between Stroop-Flanker conflict prior dPC and the task dPC is 74.42° and did not differ significantly from orthogonality ($p = 0.30$, tau = -0.038, Kendall rank correlation).

(f) Same as in (e) but for Flanker prior.

(g-j) Contribution of single neuron coding of error (g), Stroop-Simon ex-ante conflict (h), Stroop-Simon ex-post conflict (i) and Stroop-Simon prior (j) to the task-invariant population coding of these variables. Because MSIT has two conflict conditions, Simon and Flanker, task invariance was tested between Stroop and Simon or between Stroop and Flanker separately. We modelled each neuron's baseline (j), ex-ante (h) or ex-post (g, i) response using ANOVA. The main effects are a dummy variable indicating error (g) or Stroop-Simon conflict (h, i) or Stroop-Simon prior (j) and task ID (Stroop or MSIT), and the interaction term between these two. A significant interaction suggests that the coding is more prominent in one task than the other. Task-invariant neurons is defined as having a significant main effect of the variable of interest but an insignificant interaction with the task ID. Task-dependent neurons is defined as having a significant interaction term.

(g) Contribution of single neuron ex-post coding of error to task-invariant population coding of error. Of the 37 of neurons that were selected as signaling error in the ex-post epoch in either task, 68% did so in a task-invariant way (red) and 32% in a task-dependent (blue) way (pie chart). There is a strong correlation between the error t-statistic and the dPCA weight of a particular neuron (scatter plot on the left; $R = 0.72$). Comparing the mean absolute value of dPCA weights between task-invariant, task-dependent and uncategorized neurons, both the task-invariant ($p < 0.01$) and task-dependent ($p < 0.05$) neurons had significantly larger absolute weights than the uncategorized neurons.

(h) Contribution of single neuron ex-ante coding of Stroop-Simon conflict to task-invariant population coding of Stroop-Simon conflict. Of the 40 of neurons that were selected as signaling conflict in the ex-ante epoch in either task, 40% did so in a task-invariant way (red) and 60% in a task-dependent (blue) way (pie chart). There is a strong correlation between the error t-statistic and the dPCA weight of a particular neuron (scatter plot on the left; $R = 0.51$). Comparing the mean absolute value of dPCA weights between task-invariant, task-dependent and uncategorized neurons, only the task-invariant neurons had significantly larger absolute weights than the uncategorized neurons ($p < 0.01$).

(i) Contribution of single neuron ex-post coding of Stroop-Simon conflict to task-invariant population coding of Stroop-Simon conflict. Of the 46 of neurons that were selected as signaling conflict in the ex-post epoch, 46% did so in a task-invariant way (red) and 54% in a task-dependent (blue) way (pie chart). There is a strong correlation between the error t-statistic and the dPCA weight of a particular neuron (scatter plot on the left; R = 0.51). Comparing the mean absolute value of dPCA weights between task-invariant, task-dependent and uncategorized neurons, only the task-invariant neurons had significantly larger absolute weights than the uncategorized neurons (p < 0.01).
(j) Contribution of single neuron baseline coding of Stroop-Simon conflict prior to task-invariant population coding of Stroop-Simon conflict prior. Of the 75 of neurons that were selected as signaling prior in the baseline in either task, 49% did so in a task-invariant way (red) and 51% in a task-dependent (blue) way (pie chart). There is a strong correlation between the error t-statistic and the dPCA weight of a particular neuron (scatter plot on the left; R = 0.51). Comparing the mean absolute value of dPCA weights between task-invariant, task-dependent and uncategorized neurons, both the task-invariant (p < 0.001) and task-dependent (p < 0.001) neurons had significantly larger absolute weights than the uncategorized neurons.

*p < 0.05, ** p < 0.01, *** p <= 0.001, n.s., not significant (p > 0.05).

Table S1. Number of sessions and neurons recorded
Summary of number of neurons recorded in each subject. For some subjects, both the Stroop task and MSIT were performed.

Patients ID	Sex	Age	Stroop			MSIT		
			# sessions	dACC	pSMA	# sessions	dACC	pSMA
P9HMH	M	55	1	9	0	NA	NA	NA
P11HMH	M	16	2	26	0	NA	NA	NA
P14HMH	M	31	2	5	0	NA	NA	NA
P16HMH	F	34	2	22	0	NA	NA	NA
P19HMH	M	34	1	13	0	NA	NA	NA
P21HMH	M	20	2	8	0	NA	NA	NA
'P31HMH'	M	30	1	3	0	NA	NA	NA
'P41HMH'	M	19	1	2	0	NA	NA	NA
'P42HMH'	M	29	1	2	0	NA	NA	NA
'P24CS'	F	47	2	17	46	NA	NA	NA
'P25CS'	F	36	2	32	0	NA	NA	NA
'P26CS'	F	36	1	16	0	NA	NA	NA
'P27CS'	M	45	1	2	2	NA	NA	NA
'P29CS'	M	19	1	9	9	NA	NA	NA
'P31CS'	M	31	2	38	9	NA	NA	NA
'P32CS'	M	19	1	0	5	NA	NA	NA
'P33CS'	F	44	4	66	25	NA	NA	NA
'P34CS'	M	70	5	26	0	1	4	0
'P35CS'	M	63	6	1	47	1	0	9
'P36CS'	M	45	6	8	64	NA	NA	NA
'P37CS'	F	33	11	107	40	5	64	17
'P39CS'	M	26	6	21	96	NA	NA	NA
'P40CS'	M	25	3	7	25	3	20	34
'P42CS'	F	25	5	83	60	8	45	97
'P47CS'	M	33	2	0	18	NA	NA	NA
'P48CS'	F	32	1	20	21	NA	NA	NA
'P44CS'	F	53	2	20	42	2	17	36
'P49CS'	F	24	NA	NA	NA	1	0	4
'P51CS'	M	17	NA	NA	NA	9	107	17
'P55CS'	F	43	NA	NA	NA	4	14	67
'P56CS'	M	48	3	4	15	NA	NA	NA
'P60CS'	M	67	NA	NA	NA	2	29	49
'P61CS'	F	52	4	7	79	4	7	77
'P71CS'	M	40	1	19	4	1	19	5

Table S2. Model comparisons for RT

Table S2

BIC of RT

	RT Tuned	no RT tuning	RL
MSIT	-686.3	-337.8	-450.9
Stroop	-1412.1	-904.1	-972.5

	RT Tuned	Prev conflict
MSIT	-805.7	-440.1
Stroop	-1733.6	-1435.1

Model comparison for RT using BIC. To test whether our RT-tuned Bayesian model explains variance in RT better than other models, we used linear mixed-effect models that takes into account subject variability (details of the model see **Methods**) and computed BIC for these models. The conflict prior is entered as a main fixed-effect and also as a by-session random effect. Here, conflict priors generated by four models are considered: "RT tuned", Bayesian conflict learning model with DDM hyperparameters and thus the conflict prior is tuned by RT. "No RT tuning", Bayesian conflict learning model without incorporating DDM likelihood for RT. "RL", a reinforcement learning model where the conflict probability is modelled as a "value" function and updated trial-by-trial by a simple update rule. "Prev conflict", a dummy variable indicating previous trial conflict. These linear mixed-effect models all have the same number of free parameters. A separate comparison was done between the RT tuned model with the model that uses the previous conflict (sub-table on the right) because the number of trials must be kept the same for the comparison and the "prev conflict" model did not consider the first trial for each session (there was no "prev conflict" in that case).

Table S3 Model comparison for trial congruency

Table S3

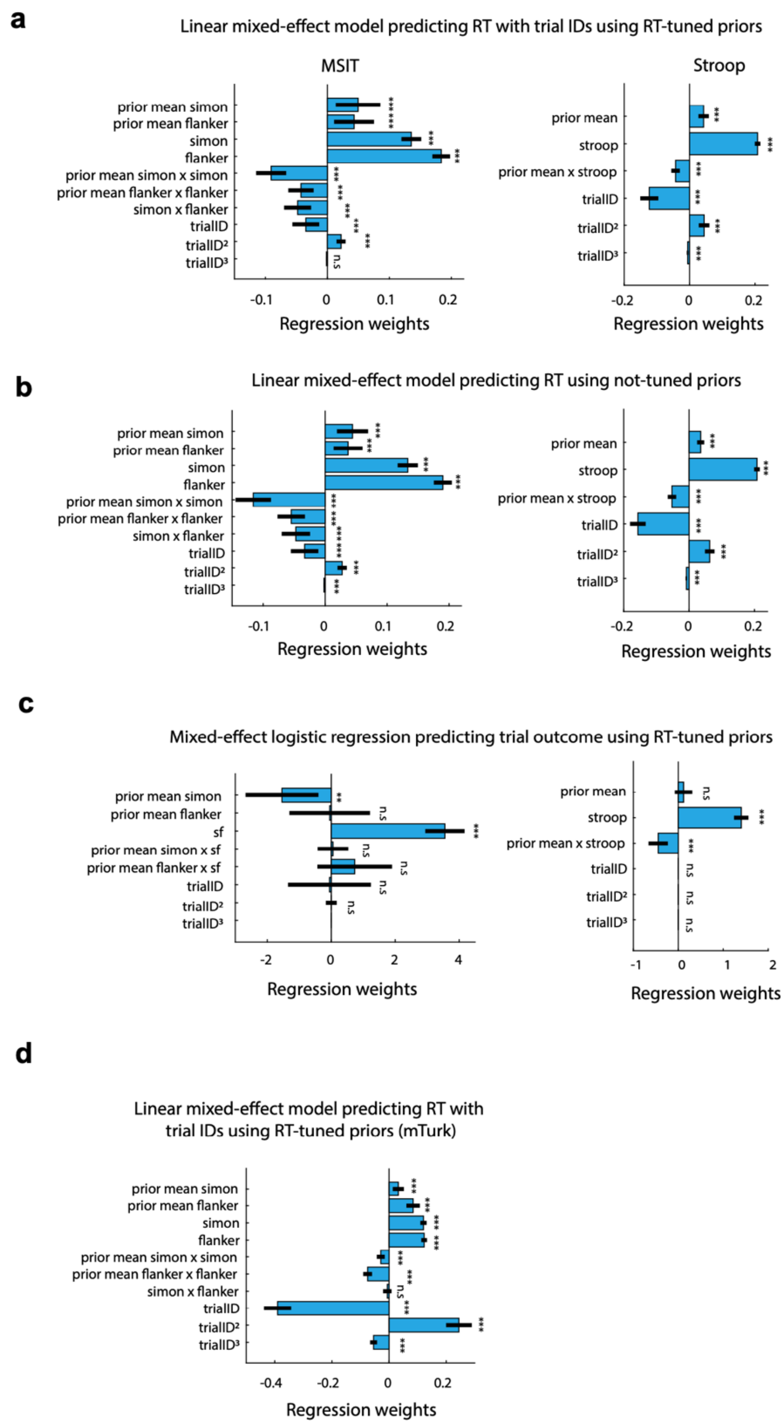
BIC of conflict

	RT Tuned	no RT tuning	RL	constant prior
MSIT	26517	26674	26958	28271
Stroop	23807	24039	24779	26498

Model comparison for trial congruency using BIC. We used Bernoulli likelihood when computing BIC for the conflict sequence. Note that the number of fitted parameters for Bayesian models is zero, for the "RL" model is one (learning rate), and for "constant prior" model is one (the constant prior). BIC penalizes free parameters.

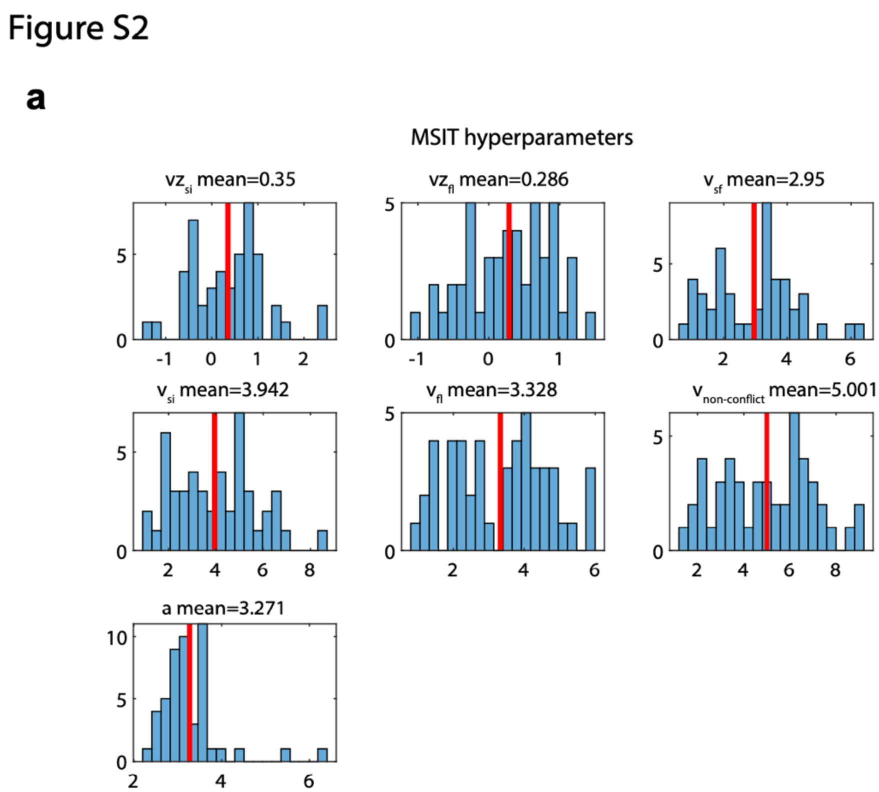
Figure S1 Behavioral models. Related to Figure 1.

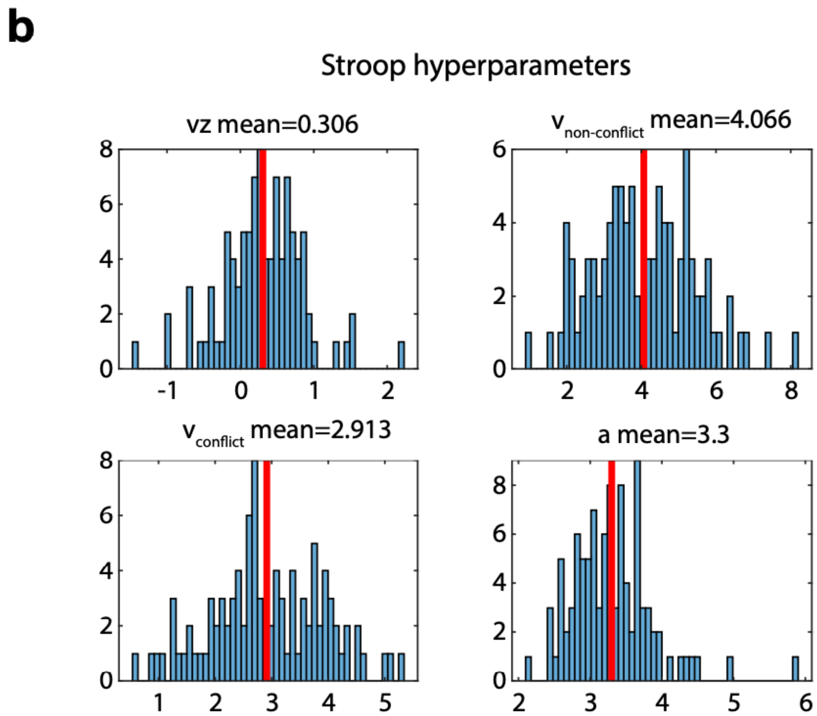
Figure S1



Statistical significance of regressors is determined by comparing the full model and a reduced model with a particular regressor removed, using a likelihood ratio test.
(a) Linear mixed-effect model for RT that incorporates trial ID regressors for MSIT (left) and Stroop (right). Conflict priors used are from the Bayesian online learning models with RT tuning. We added the first, second and third -order trial ID regressors to model putative practice effects. The main effects of conflict priors, conflict, and their interaction are all significant even in the presence of trial ID regressors, suggesting these regressors capture behavioral effect that do not depend on trial ID.
(b) Same as (a), but for the Bayesian online learning models without RT tuning. Thus, in this instance, conflict prior is estimated based on conflict sequence alone. The main effects of conflict priors (not tuned by RT), conflict, and their interaction are all significant in the presence of trial ID regressors. Therefore, RT tuning improves conflict prior (see (a) and Table S2, S3), but this is not required.
(c) Mixed-effect logistic regression for predicting trial outcome (error or correct) for MSIT (left) and Stroop (right). Conflict priors used are from the Bayesian online learning models with RT tuning. For MSIT, we consider only "sf" trials for conflict trials, on which most of errors occur, and non-conflict trials. Conflict prior reduces error likelihood in both MSIT (significant main effect, $p = 0.009$) and Stroop (significant interaction term), $p < 0.001$).
(d) Linear mixed-effect model for RT that incorporates trial ID regressors for MSIT. Data were collected from online participants using Amazon Mechanical Turk.

Figure S2. DDM hyperparameters used in Bayesian conflict learning models. Related to Figure 1.





(a) Hyperparameters used in the Bayesian conflict learning model for MSIT. vz_{si} and vz_{fl} are coefficients scaling Simon and Flanker prior. v_{sf}, v_{si}, v_{fl}, $v_{non-conflict}$ are base drift rates in both Simon and Flanker present ("sf"), Simon-only ("si"), Flanker-only ("fl"), non-conflict trials. a is the boundary separation. The effective drift rate was the sum of the base drift rate and the scaled conflict prior. The base drift rates were significantly different from each other ($p < 0.001$, ANOVA). Post-hoc pairwise testing from a multiple comparison test determined that v_{si} did not differ significantly from v_{fl}; v_{sf} were significantly larger than either v_{si} or v_{fl}; both v_{si} and v_{fl} were significantly larger than $v_{non-conflict}$. These

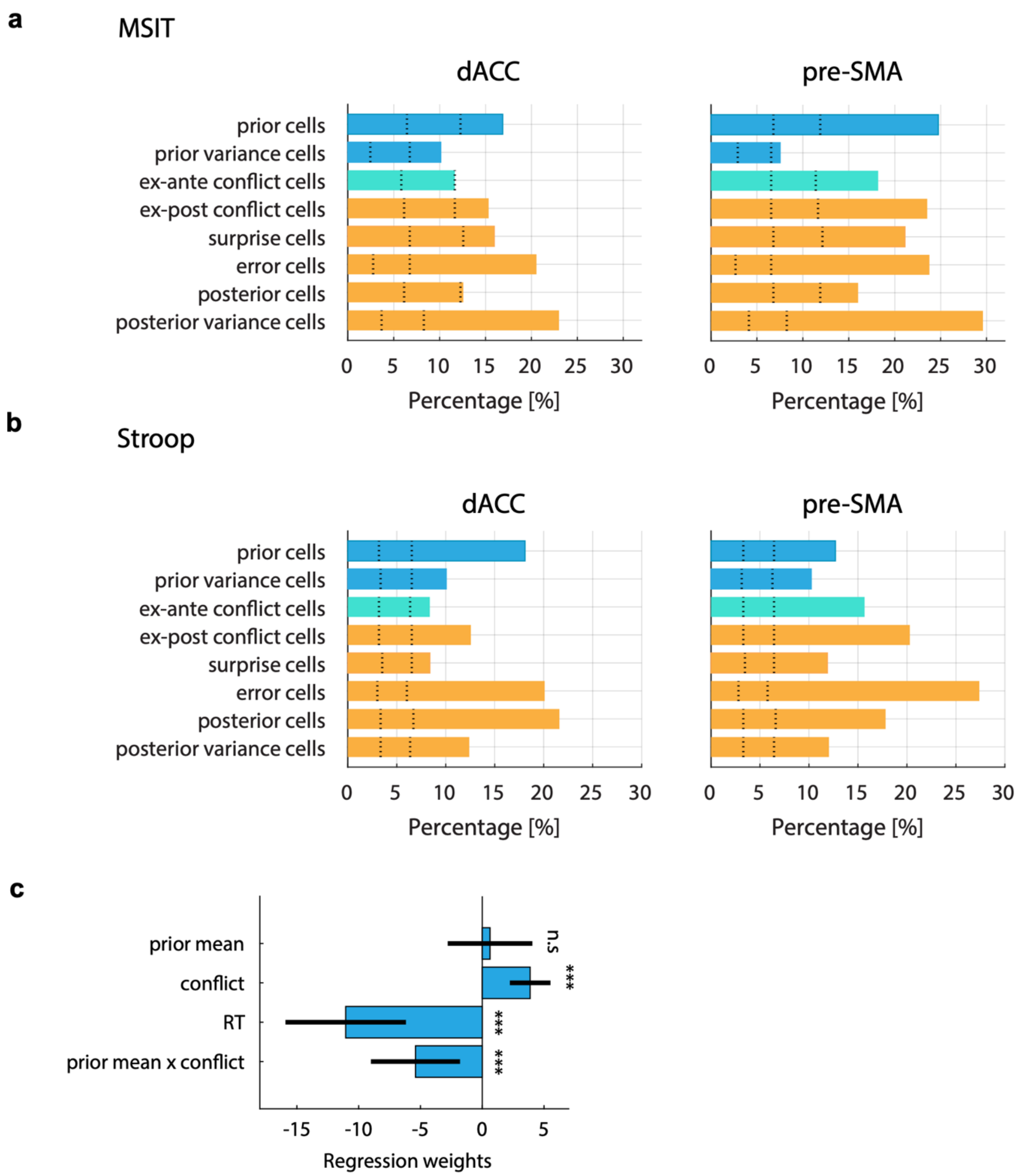
(b) Hyperparameters used in the Bayesian conflict learning model for Stroop. vz is the coefficient scaling Stroop prior. $v_{conflict}$, $v_{non-conflict}$ are base drift rates in conflict and non-conflict trials. a is the boundary separation. $v_{non-conflict}$ are significantly larger than $v_{conflict}$ across sessions ($p < 0.001$, t test).

Hyperparameters are used in the DDM likelihood function for tuning the prior estimation process using an expectation-maximization algorithm.

*p < 0.05, ** p < 0.01, *** p < 0.001, n.s., not significant (p > 0.05 or not significant determined using FDR).

Figure S3. Neuronal selection by areas and ERP analysis. Related to Figure 3.

Figure S3

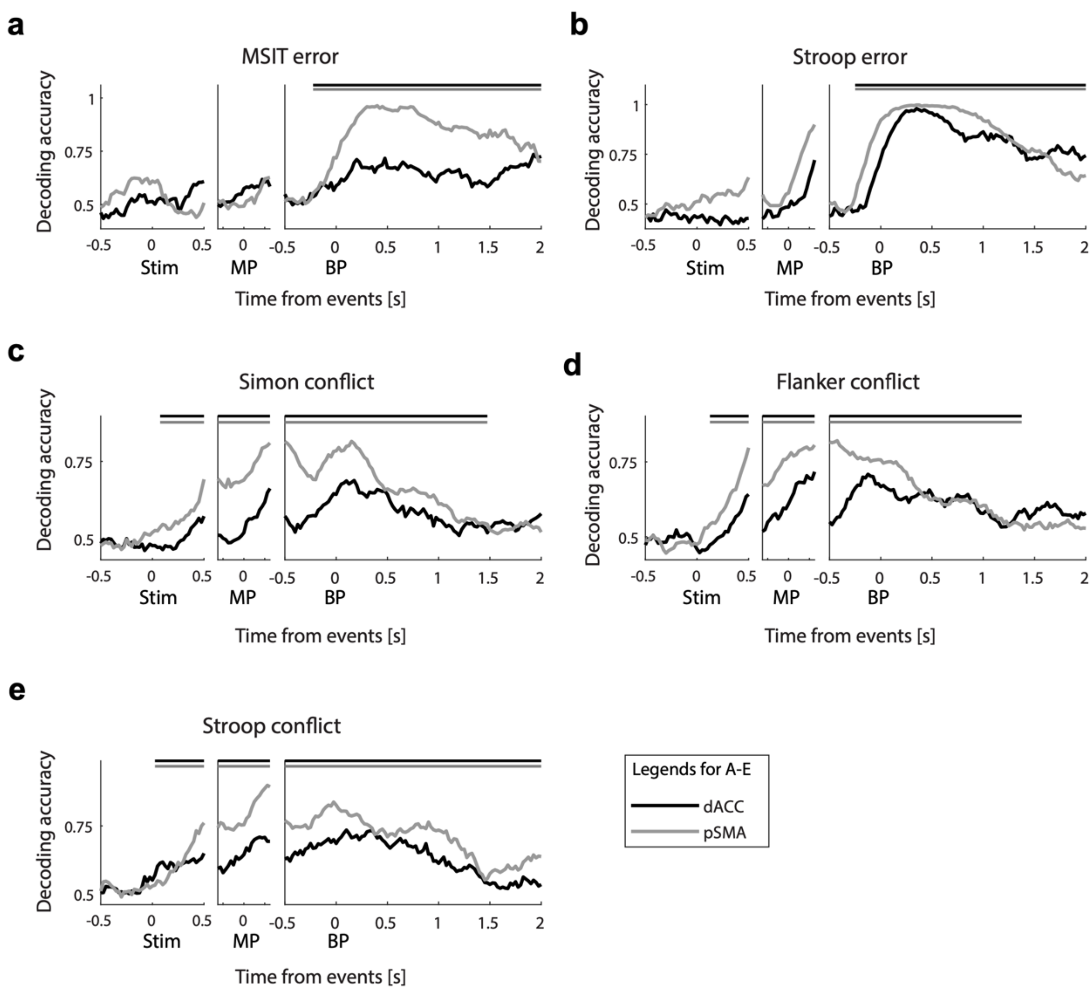


(a) Percentages of significant neurons in both dACC (left) and pre-SMA (right) in MSIT.
(b) Percentages of significant neurons in both dACC (left) and pre-SMA (right) in Stroop.
Dotted lines represent 2.5th and 97.5th percentiles of the null distribution obtained from permutation. For all groups shown, $p < 0.001$. Patterns of neuronal selection are similar between dACC and pre-SMA.
(c) Linear mixed-effect model for CRP amplitude in the Stroop task. Conflict priors used are from the Bayesian online learning models with RT tuning. The main effects of conflict, RT and the

interaction between prior and conflict were all significant. The main effect of conflict prior was not significant. Statistical significance was determined by a likelihood ratio test (comparing between the full model and the reduced models with regressors of interest removed).

Figure S4. Population decoding of error and conflict by areas. Related to Figure 5.

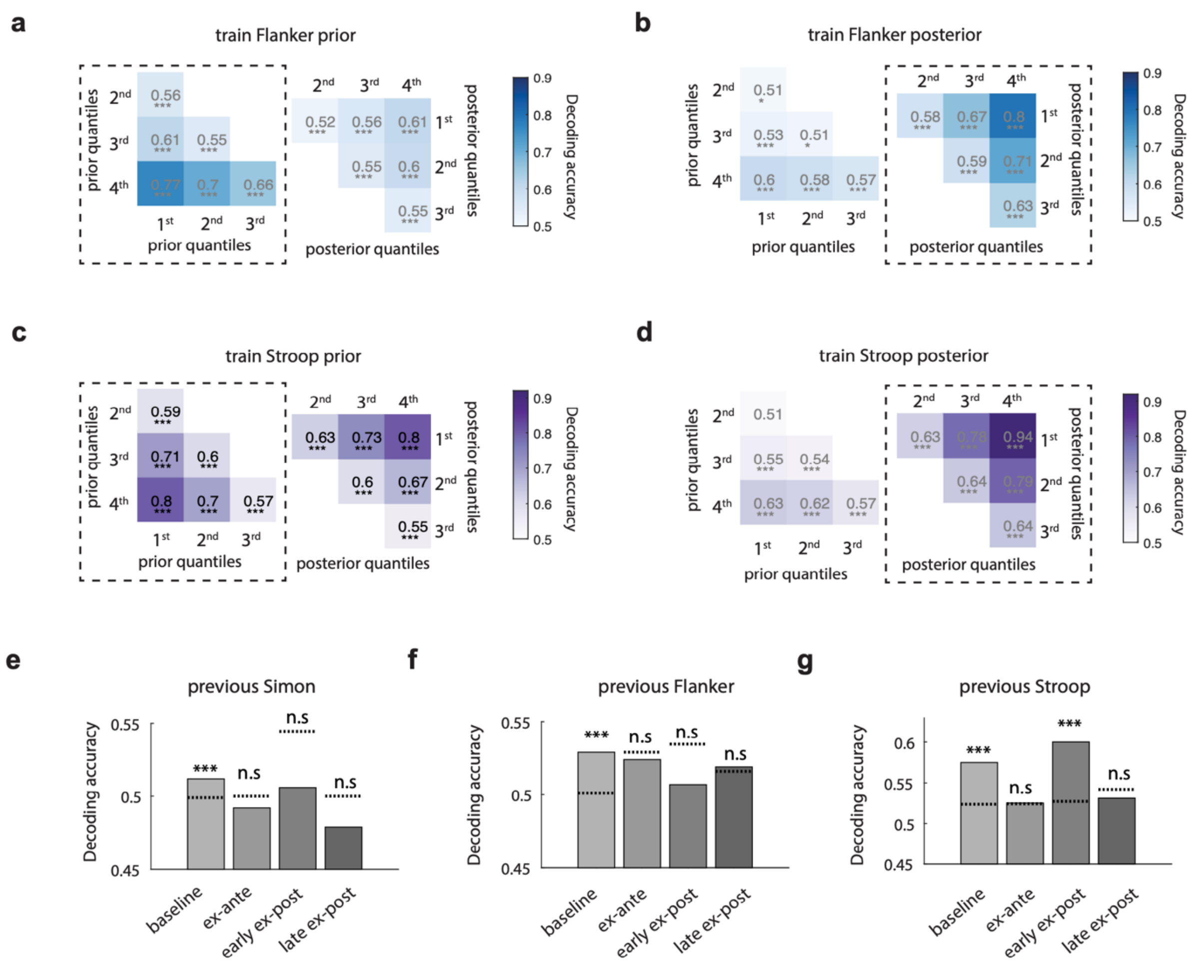
Figure S4



(a-e) Population decoding accuracy for MSIT error (a), Stroop error (b), Simon conflict (c), Flanker conflict (d), Stroop conflict (e). For (a-e), black traces are from dACC data and grey traces are from pre-SMA data. Horizontal bars at the top demarcate significant cluster as determined by the cluster-based permutation test ($p < 0.05$). Overall dynamics are similar between dACC and pre-SMA, though the decoding accuracy on average is lower in the former.

Figure S5. Population decoding of prior/posterior (Flanker and Stroop) and past-trial conflict. Related to Figure 5.

Figure S5



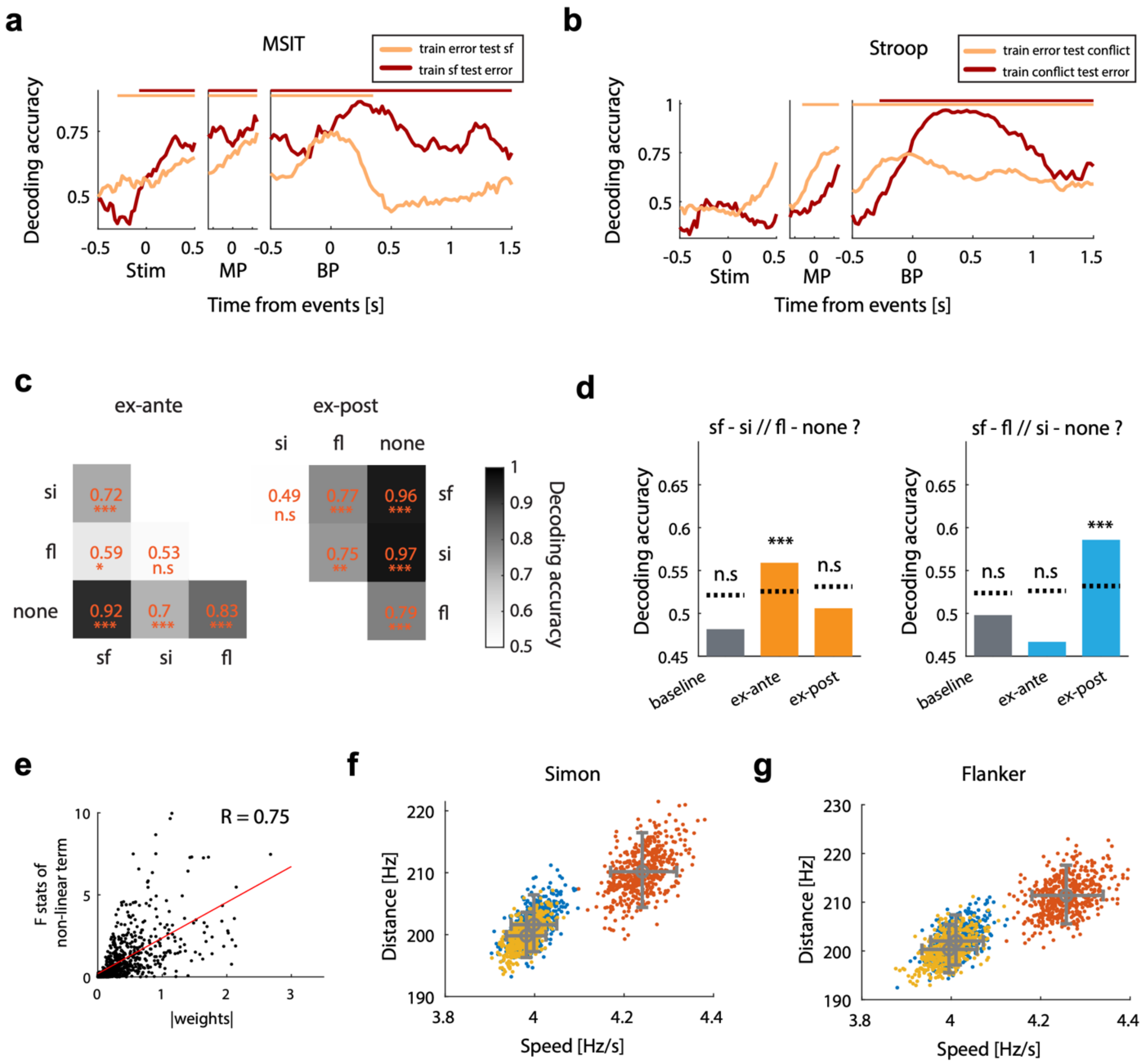
(a-d) Decoding accuracy for classifying between pairs of Flanker (a-b) and Stroop (c-d) prior or posterior quartiles. Color bars show decoding accuracy. Dotted frames mark the within-time decoding results. Decoders that classify prior quartiles are trained using baseline spike counts (1.5s before stimulus onset), whereas decoders that classify posterior quartiles are trained using ex-post spike counts (2s after button press). To test temporal generalization of these decoders, prior-trained decoders are tested with posterior data and labels, and vice versa. Dashed boxes represent within-time decoding.

(e-g) Population decoding of Simon (e), Flanker (f), Stroop (g) on the immediately preceding trial in different epochs. Dotted lines show 97.5% percentile from the null distribution (permutation). During baseline, there is significant coding of past trial conflict as expected from the persistence of ex-post conflict signals. Coding of the past trial conflict was non-significant during all other epochs except for past trial Stroop conflict in the early ex-post epochs, suggesting that this information in our experimental setup was likely not reliable for cognitive control.

*p < 0.05, ** p < 0.01, *** p < 0.001, n.s., not significant (p > 0.05 or not significant determined using FDR).

Figure S6. Within-task state space analyses. Related to Figure 6.

Figure S6



(a) A common population coding dimension for error and conflict in MSIT. This coding dimension is extracted using dPCA, using an error contrast (error "sf" trials vs. "sf" trials) and a conflict contrast ("sf" trials vs. no conflict). "sf" trials are split into two non-overlapping groups for this. Plot show the decoding accuracy of both sf (apricot) vs no-conflict trials, and error "sf" vs. "sf" trials (out-of-sample). Horizontal bars at the top demarcate significant clusters, as determined by the cluster-based permutation test ($p < 0.05$).

(b) A common population coding dimension for both error and conflict in Stroop. This coding dimension is extracted using dPCA, using an error contrast (error conflict trials vs. conflict trials) and a conflict contrast (conflict trials vs. no conflict) and by marginalizing out the contrast dimension. Conflict trials are split into two non-overlapping groups for this. Plot show the decoding accuracy of both sf (apricot) vs no-conflict trials, and error "sf" vs. "sf" trials (out-of-sample). Horizontal bars at the top demarcate significant clusters, as determined by the cluster-based permutation test ($p < 0.05$).

(c) Decoding accuracy of pairwise classification of conflict conditions after RT was equalized across conditions. Trials were selected such that RTs on si, fl, sf and non-conflict trials were equalized ($p > 0.1$, t test). Training data and left-out testing data from the four conflict conditions are projected to the population vector flanked by averages of non-conflict trials and sf trials, shown as the dotted line in Figure 6a. Color code represents decoding accuracy. This coding dimension separates the four conflict conditions well.

(d) Testing compositionality of conflict representation with condition generalization of decoding on error trials. We tested the condition generalization using raw spike count data from the ex-ante and ex-post (1s after button presses) epochs and from the baseline as a control without dimensionality reduction. We only used data on error trials for this analysis. Data from the means connected by one of the blue edges in Figure 6a were used to train a decoder, which was then tested with left-out data from the means connected by the opposite blue edge, and vice versa. Blue bars show decoding accuracy. Same was also tested with data connected by the orange edges as shown in Figure 6a. Orange bars show decoding accuracy. Decoding accuracy were reduced on error trials compared to on correct trials (compare with Figure 6d). Dotted lines show 97.5th percentile of the null distribution from permutation.

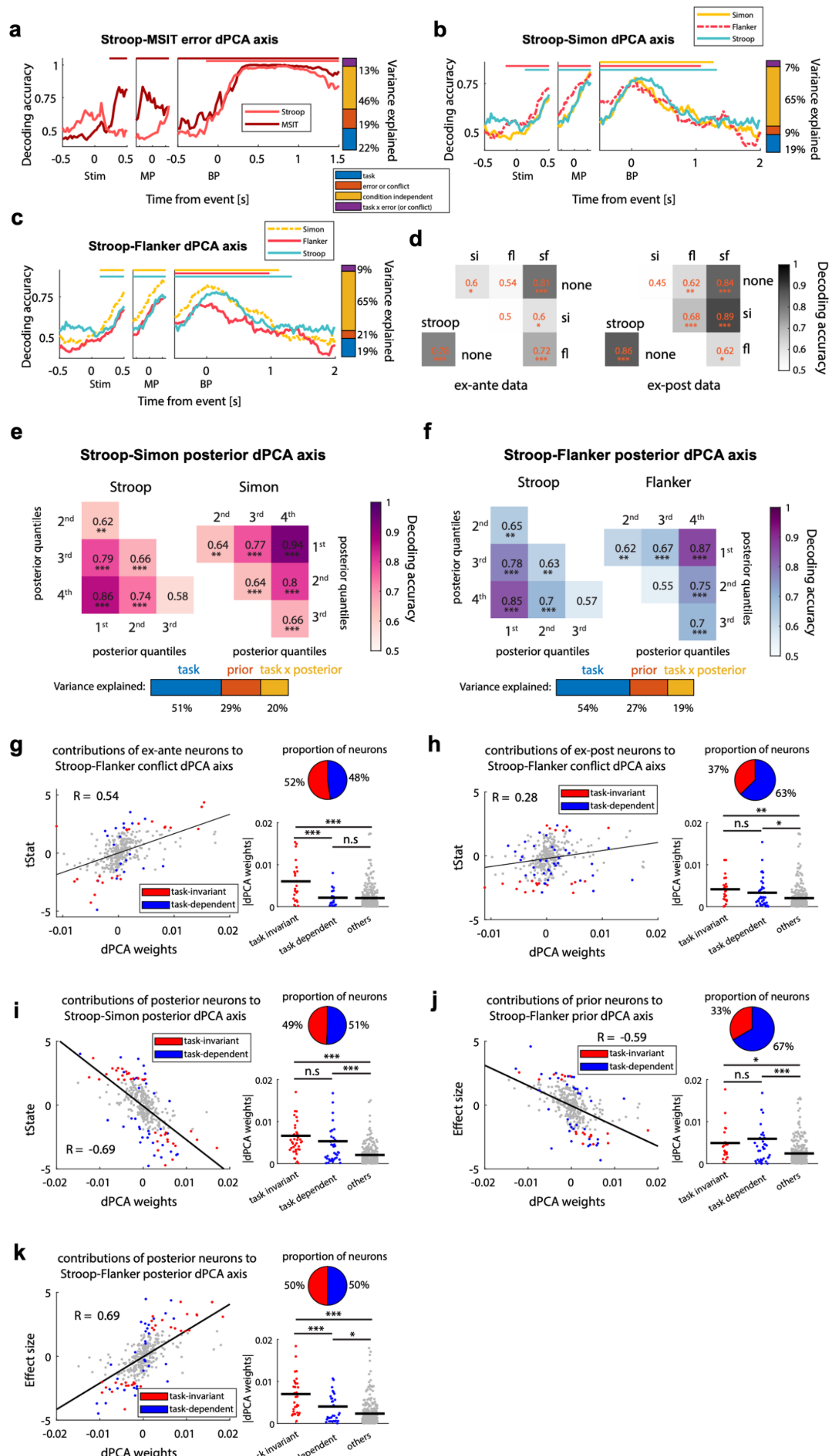
(e) Single neuron with nonlinear coding of Simon and Flanker conflict contribute to deviation of conflict representation from perfect linearity. Data used here are from the ex-ante epoch. Nonlinear coding of conflict by a single neuron is measured by the F statistic of the interaction term between Simon and Flanker conflict in an ANOVA model with spike counts as the dependent variable. Each neuron's contribution to the deviation from linear additivity in the high dimensional neural space is quantified by the weight of the difference vector between "sf" and "si + fl". Scatter plot shows the relation between these two measures. Red line shows the linear fit.

(f) Distance between trajectories and average speed computed from trials grouped by quartiles of Simon conflict prior in the baseline (blue) and the ex-ante (orange) epoch, and trials grouped by Simon conflict posterior in the ex-post epoch (yellow). Trajectories are visualized in Figure 6f. The state space speed stays low during baseline, increases significantly during the ex-ante epoch and decreases back to a value similar to that during the baseline. Distance between trajectories is stable across time.

(g) Same as in (f) but for Flanker conflict prior.

Figure S7. Domain-general representation of performance monitoring signals. Related to Figure 7.

Figure S7


a
Stroop-MSIT error dPCA axis
Decoding accuracy
Stroop
MSIT
Stim
MP
BP
Time from event [s]
13%
46%
19%
22%
Variance explained
task
error or conflict
condition independent
task x error (or conflict)
b
Stroop-Simon dPCA axis
Simon
Flanker
Stroop
7%
65%
9%
19%
c
Stroop-Flanker dPCA axis
9%
65%
21%
19%
d
si
fl
sf
none
stroop
ex-ante data
ex-post data
Decoding accuracy
e
Stroop-Simon posterior dPCA axis
Stroop
Simon
posterior quantiles
Variance explained:
task
prior
task x posterior
51%
29%
20%
f
Stroop-Flanker posterior dPCA axis
Stroop
Flanker
54%
27%
19%
g
contributions of ex-ante neurons to Stroop-Flanker conflict dPCA aixs
proportion of neurons
R = 0.54
tStat
dPCA weights
task-invariant
task-dependent
|dPCA weights|
task invariant
task dependent
others
h
contributions of ex-post neurons to Stroop-Flanker conflict dPCA aixs
R = 0.28
i
contributions of posterior neurons to Stroop-Simon posterior dPCA axis
R = -0.69
tState
j
contributions of prior neurons to Stroop-Flanker prior dPCA axis
R = -0.59
Effect size
k
contributions of posterior neurons to Stroop-Flanker posterior dPCA axis
R = 0.69

(a) Task-invariant decoding of errors in both MSIT (salmon) and Stroop (crimson) after RTs were equalized across conditions. Specifically, trials were selected for Stroop such that the RTs on error conflict trials did not differ significantly those from correct conflict trials (p > 0.1, t-test). For MSIT, trials were selected such that RTs on error "sf" trials and correct "sf" trials did not differ significantly (p > 0.1, t test). The task-invariant coding dimension is extracted using dPCA that marginalizes out task information and time. This dPCA coding dimension is extracted from the error contrast in Stroop (error conflict vs. correct conflict trials) and the error contrast in MSIT (error "sf" trials and correct "sf" trials). This controls for trial conflict and isolates effects related only to error. Left, Accuracy for decoding errors as a function of time. Bar on the right shows the variance explained by the different dPCA components (color code see figure legend).

(b-c) Task-invariant decoding of conflict in both MSIT (Simon, yellow; Flanker red) and Stroop (green) after RTs were equalized across conditions. Because MSIT has two conflict conditions, Simon and Flanker, task invariance was investigated using Stroop/Simon and Stroop/Flanker conflicts separately. Specifically, trials were selected for Stroop such that RTs on conflict and non-conflict trials did not differ significantly (p > 0.1, t test). For MSIT, trials were selected such that RTs on Simon and non-Simon trials did not differ significantly (p > 0.1) and those on Flanker and non-Flanker trials did no differ significantly (p > 0.1). This dPCA coding dimension is extracted from conflict and non-conflict trials in Stroop and either from Simon and non-Simon trials (b) or from Flanker and non-Flanker trials (c), by marginalizing out task information and time. Left-out conflict trials and non-conflict trials in Stroop, and left-out Simon, non-Simon, Flanker, non-Flanker trials in MSIT are projected and classified by this coding dimension. Left, decoding accuracy of conflict as a function of time. The bar on the right shows variance explained by the different dPCA components (color code see figure legend).

(d) Testing separability of conflict conditions in Stroop and MSIT using data from the ex-ante (left) and ex-post epochs (right) after RTs were equalize across conditions. Here, trials were selected for Stroop such that RTs on conflict and non-conflict trials did not differ significantly (p > 0.1, t test). For MSIT, trials were selected such that RTs on si, fl,sf, and non-conflict trials did not differ with each other significantly (p > 0.1, t test). The dPCA coding dimension used in this analysis is extracted by using conflict and non-conflict trials in Stroop and sf and non-conflict trials in MSIT by marginalizing out task information. Because data from ROIs are used, temporal information is already marginalized out before being used by the dPCA algorithm. This coding dimension support classification of 75% of pairs of MSIT conflict conditions (upper triangle matrices) as well as Stroop conflict (left corner). Color coding represents decoding accuracy, orange numbers indicate the numerical values of decoding accuracy of that pair of conflict conditions.

(e) Task-invariant decoding of all pairs of conflict posterior quartiles in Stroop (lower triangle matrix) and Simon (upper triangle matrix). The dPCA coding dimension here is extracted using the Stroop conflict posterior contrast (the 1st vs. 4th quartiles of Stroop conflict posterior) and the Simon conflict posterior contrast (the 1st vs. 4th quartiles of Simon conflict posterior), marginalizing out task information. Color code represents decoding accuracy. Bar at the bottom shows variance explained of dPCA components (for decoding, the component labelled as "posterior" is used).

(f) Same as in (e) but for Flanker posterior.

(g-k) Contribution of single neuron coding of Stroop-Flanker ex-ante conflict (g), Stroop-Flanker ex-post conflict (h), Stroop-Simon posterior (i), Stroop-Flanker prior (j) and Stroop-Flanker posterior (k) to the task-invariant population coding of these variables. Because MSIT has two conflict conditions, Simon and Flanker, task invariance was tested between Stroop and Simon or between Stroop and Flanker separately. We modelled each neuron's baseline (j), ex-ante (g) or ex-post (h,i,k) response using linear regression. The main effects are a dummy variable indicating Stroop-Flanker conflict (g,h) or Stroop-Simon posterior (i) or Stroop-Flanker prior (j) or Simon-Flanker posterior (k) and task ID (Stroop or MSIT), and the interaction term between these two. A significant interaction suggests that the coding is more prominent in one task than the other. Task-invariant neurons is defined as having a significant main effect of the variable of interest but an insignificant interaction with the task ID. Task-dependent neurons is defined as having a significant interaction term.

(g) Contribution of single neuron ex-ante coding of Stroop-Flanker conflict to task-invariant population coding of Stroop-Flanker conflict. Of the 42 of neurons that were selected as signaling conflict in the ex-ante epoch in either task, 52% did so in a task-invariant way (red) and 48% in a task-dependent (blue) way (pie chart). There is a strong correlation between the error t-statistic and the dPCA weight of a particular neuron (scatter plot on the left; $r = 0.54$). Comparing the mean absolute value of dPCA weights between task-invariant, task-dependent and uncategorized neurons, the task-invariant neurons had significantly larger absolute weights than the task-dependent neurons ($p < 0.001$) and uncategorized neurons ($p < 0.001$).

(h) Contribution of single neuron ex-post coding of Stroop-Flanker conflict to task-invariant population coding of Stroop-Flanker conflict. Of the 59 of neurons that were selected as signaling conflict in the ex-post epoch in either task, 37% did so in a task-invariant way (red) and 67% in a task-dependent (blue) way (pie chart). There is a strong correlation between the error t-statistic and the dPCA weight of a particular neuron (scatter plot on the left; $r = 0.28$). Comparing the mean absolute value of dPCA weights between task-invariant, task-dependent and uncategorized neurons, both the task-invariant ($p < 0.01$) and task-dependent ($p < 0.05$) neurons had significantly larger absolute weights than the uncategorized neurons.

(i) Contribution of single neuron ex-post coding of Stroop-Simon conflict posterior to task-invariant population coding of Stroop-Simon conflict prior. Of the 79 of neurons that were selected as signaling posterior in the ex-post epoch in either task, 49% did so in a task-invariant way (red) and 51% in a task-dependent (blue) way (pie chart). There is a strong correlation between the error t-statistic and the dPCA weight of a particular neuron (scatter plot on the left; $R = 0.51$). Comparing the mean absolute value of dPCA weights between task-invariant, task-dependent and uncategorized neurons, both the task-invariant ($p < 0.001$) and task-dependent ($p < 0.001$) neurons had significantly larger absolute weights than the uncategorized neurons.

(j) Contribution of single neuron baseline coding of Stroop-Flanker conflict prior to task-invariant population coding of Stroop-Flanker conflict prior. Of the 58 of neurons that were selected as signaling prior in the baseline in either task, 33% did so in a task-invariant way (red) and 67% in a task-dependent (blue) way (pie chart). There is a strong correlation between the error t-statistic and the dPCA weight of a particular neuron (scatter plot on the left; $r = -0.59$). Comparing the mean absolute value of dPCA weights between task-invariant, task-dependent and uncategorized neurons, both the task-invariant ($p < 0.05$) and task-dependent ($p < 0.001$) neurons had significantly larger absolute weights than the uncategorized neurons.

(k) Contribution of single neuron ex-post coding of Stroop-Flanker conflict posterior to task-invariant population coding of Stroop-Flanker conflict posterior. Of the 66 of neurons that were selected as signaling posterior in the ex-post epoch in either task, 50% did so in a task-invariant way (red) and 50% in a task-dependent (blue) way (pie chart). There is a strong correlation between the error t-statistic and the dPCA weight of a particular neuron (scatter plot on the left; r = 0.69). Comparing the mean absolute value of dPCA weights between task-invariant, task-dependent and uncategorized neurons, both the task-invariant ($p < 0.001$) and task-dependent ($p < 0.05$) neurons had significantly larger absolute weights than the uncategorized neurons, and the task-invariant had significantly larger values than the task-dependent neurons ($p < 0.001$).

Methods

Tasks

Subjects performed a speeded version of the Stroop and Multi-Source Interference (MSIT) tasks. For the Stroop task, subjects were shown a series of randomly intermixed color words ("red", "green", "blue") printed in either red, green, or blue color (see Figure 1a). Subjects were instructed to name the color the word stimulus was printed in while ignoring its meaning and to do so as quickly and accurately as possible. For the MSIT task, subjects were shown an array of three numbers (0,1,2,3), out of which two were the same and the third of which was different (target). Subjects were instructed to press the button identical to the target number (which was unique) regardless of the position at which it was shown. For both tasks, all responses were recorded as button presses using an external response box (RB-740, Cedrus Corp., San Pedro, CA). For both tasks, the stimulus disappeared immediately when a button was pressed and was followed by a blank screen for 1s, followed by a feedback screen, which was shown for 1s. Subject were given three types of feedback: correct, incorrect or "too slow". 10-15% of trials were rated as "too slow" based on an adaptive response threshold (see (Fu et al., 2019) for details), which we used to emphasize the need to respond quickly and thereby resulting in a sufficiently large error rate (~10% of trials). The inter-trial interval was sampled randomly from a uniform distribution between 1.5s to 2s. Trial sequences were pseudo-randomized and designed to avoid back-to-back repetitions of the same stimulus. The proportion of conflict trials in the Stroop task was 30-40%; For MSIT, the proportions of Simon only ("si"), Flanker only ("fl"), Simon and Flanker coincident ("sf") trials are 15%, 15%, and 30%, respectively (the remaining 40% of trials have no conflict). The tasks were implemented in MATLAB (The Mathworks, Inc., Natick, MA) using Psychtoolbox-3 (Brainard, 1997). The two tasks were performed in sequence, i.e., subjects finished blocks of one task first and then moved on to blocks of the other task. The order of task performed was randomized across experimental sessions.

Behavioral controls

As a control, we additionally collected behavioral data from N = 51 normal control subjects (24 females; age mean$\pm$sd: 44$\pm$11) using the Amazon mTurk platform. We implemented the MSIT task as described above using the jsPsych toolbox (de Leeuw, 2015). These behavioral data were analyzed using the same methods as documented below. These control subjects exhibited a robust conflict prior effect like the patients (see Figure S1d).

Subjects

34 patients (see Table S1 for age and gender) who were evaluated for possible surgical treatment of their focal 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. All research protocols were approved by the institutional review boards of Cedars-Sinai Medical Center, Huntington Memorial Hospital, and California Institute of Technology.

Electrophysiological recordings

We analyzed data from up to 4 electrodes in each subject (bilateral dACC and pre-SMA) in this paper. For each depth electrode, there are eight microwires with high impedance microwires at the tip, and eight macro contacts with low impedance along the shaft (AdTech Medical Inc.). Data from all microwires and the most medial macro contact (which is placed within dACC or pre-SMA) are analyzed in this paper. For recordings from microwires, the sampling rate was 32-40kHz and the raw signal was acquired broadband (0.01Hz-9kHz). One microwire on each depth electrode was designated as a local reference wire. For intracranial EEG recordings done with macro contacts, the sampling rate was 2kHz (ATLAS, Neuralynx, Inc., Bozman, MT).

Electrode localization

Electrodes were localized using a combination of a pre-operative MRI and a postoperative MRI/CT using standard procedures we described elsewhere (Fu et al., 2019; Minxha et al., 2017). 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.

Spike detection and sorting

We filtered the raw broadband 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). Sorting quality is evaluated using the same metrics reported in (Fu et al., 2019) and only well-isolated single units are included in this paper. Channels with interictal epileptic events were excluded.

Quantification and statistical analyses

Behavioral modelling and analyses

We developed a series of Bayesian conflict learning models to infer subjects' internal estimate of conflict probability (details see below). For this analysis, we concatenated all blocks of an experiments done in a single session. For trials with unusually long RTs (> 3 sd from the mean of the whole experiment), we replaced the outlier's RT with the average RT computed from the neighboring 6 trials. We estimate the Bayesian model parameters using all trials but excluded error trials (after fitting) for analyses that focuses on conflict and conflict prior. We then analyzed whether the variance of RT was related to the estimated parameters using linear mixed-effect models (Aarts et al., 2014). For MSIT, the linear mixed-effect model is specified as follows (all models are represented in Wilkinson's notation):

$$\log(RT) \sim Simon\ prior * Simon\ conflict + Flanker\ prior * Flanker\ conflict + (1 + Simon\ prior + Flanker\ prior | sessionID : subjectID)$$

For Stroop, this is specified as:

$$\log(RT) \sim Stroop\ prior * Stroop\ conflict + (1 + Stroop\ prior | sessionID : subjectID)$$

Here, the fixed effects of Simon, Flanker and Stroop conflicts are dummy variables indicating whether a particular trial involves conflict (value = 1) or not (value = 0). The fixed effects for priors are obtained from our Bayesian conflict learning models as detailed below. To test if RT was affected by conflict on the immediately preceding trial, represented by $Simon\ prevConflict$, $Flanker\ prevConflict$ and $Stroop\ prevConflict$, we again constructed a linear mixed-effect models for both MSIT and Stroop. For MSIT, the model is specified as follows:

$$\log(RT) \sim Simon\ prevConflict * Simon\ conflict + Flanker\ prevConflict * Flanker\ conflict + (1 | sessionID : subjectID)$$

For Stroop, this is specified as:

$$\log(RT) \sim Stroop\ prevConflict * Stroop\ conflict + (1 | sessionID : subjectID)$$

We investigated the effect of conflict prior on the likelihood of making an error using generalized linear mixed-effect models. For MSIT, this model is given as

$$Outcome \sim Simon\ prior * SF + Flanker\ prior * SF + (1 + Simon\ prior + Flanker\ prior | sessionID : subjectID)$$

where SF is a dummy variable indicating whether the trial has both Simon and Flanker conflict (value = 1) or non-conflict (value = 0). We restricted this analysis to sf trials because most errors occurred on these trials. For Stroop, this model is given as

$$Outcome \sim Stroop\ prior * Stroop\ conflict + (1 + Stroop\ prior | sessionID : subjectID)$$

The response variable $Outcome$ is a categorical variable indicating whether the trial ended in a correct (0) or incorrect (1) response.

To determine the statistical significance of each fixed effect, we compared the full model with a reduced model where the fixed effect in question was removed using the likelihood ratio test. To determine whether RT tuning of the model (see below) is necessary for conflict prior to explain RT variance, we switched out the conflict prior with the one estimated without RT tuning and kept all other terms the same. Statistical significance determined this way was indicated by stars or "n.s" (non-significant) in Figures 1 and S1. To determine whether the conflict prior explains variance in RT and error likelihood independent of practice, which is assumed to vary with the trial number, we augmented the aforementioned mixed-effect models by including three additional trial-ID terms: $trialID, trialID^2, trialID^3$ to capture variance related to practice effects.

Bayesian conflict learning models

Our models are structurally similar to those used in several previous studies (Behrens et al., 2007; Jiang et al., 2015). Here, we briefly highlight the modifications we made to extend these existing models to model behavior in both the Stroop and MSIT tasks, the latter of which has two types of conflicts that are monitored at the same time. Our models have the following parameters (see Fig. 1c for a schematic of the model structure): 1) a flexible learning rate α, which captures the subject's belief in the rate of change in control demand in the environment (i.e., a change in conflict probability), and 2) conflict probability (q_s for Stroop conflict in the model for Stroop, q_{si} for Simon and q_{fl} for Flanker conflicts in the model for MSIT). The models utilize two types of data (both of which are only available after a trial's response has been made): 1) trial congruency o (value of 1 indicates an incongruent trial; o_s for Stroop congruency, o_{si} for Simon congruency and o_{fl} for Flanker congruency; 2) reaction time RT, assigning Bernoulli likelihood function to the former and the drift-diffusion model likelihood function to the latter (details see below). Prior work (Jiang et al., 2015) uses a Gaussian likelihood function to describe RT generation in a Bayesian learning framework similar to ours, but we argue that the use of DDM has several advantages over the Gaussian approach: 1) fewer parameters are used in the DDM, making it computationally possible to model two types of MSIT conflict at the same time; 2) the DDM parameters have physiological meaning and thus also provide a clear physiological reasoning for conflict prior to affect specific components of the decision process; 3) DDM has been widely used and validated as the generative framework to model RT during decision making (Pedersen et al., 2017; Wiecki et al., 2013).

Estimating control demand is operationalized as estimating the probability that a certain conflict (Stroop, Simon or Flanker) would occur in the block. One advantage of our models is that they estimate both the conflict probability and the rate of change in conflict probability in an online manner, i.e., the models iteratively update their current estimates after every trial with new incoming data (of that trial) using Bayes' law. This is consistent with the way human subjects perform conflict tasks while estimating the associated control demand: they perform and estimate trial-by-trial. Note that in this study the conflict probability was constant throughout the experiment (but this was not known to subjects). Nevertheless, we allow the model to infer the learning rate α online because humans demonstrate inherent bias in believing that environmental statistics are not stable (Yu and Cohen, 2008). There is therefore no fitting involved for estimating α and thus the models are not penalized for including this parameter in model comparisons. Note that we use a single α for both types of conflicts in MSIT. To simplify model estimation, we made the Markovian assumption that the current estimate of conflict probability depends only on the current trial congruency and RT, and the estimated conflict probability on the last trial, but not on the full history of past trial conflict probability (Behrens et al., 2007). The iterative estimation of conflict probability then involves combining the estimated conflict probability from the previous trial (prior), transition functions capturing how the current estimate will change from the previous one (the probability of current estimate conditional on previous estimate) and the likelihood function.

The model starts with a transition function for the learning rate α:

$$p(\alpha_{i+1}|\alpha_i) = k\delta(\alpha_{i+1} - \alpha_i) + 1 - k$$

This formulation assumes that the learning rate has a probability k of having the same distribution as that of the preceding trial but with a probability 1-k of switching to a uniform distribution (over all possible α), because the learning rate is largely stable across time. The transition function for conflict probability concerning the transition from the current estimate to a future estimate is computed in two steps. Here, we refer to the current-trial estimate of conflict probability for Stroop, Simon or Flanker generically as q_i, to which we assigned a uniform prior. The transition function is thus denoted as $p(q_{i+1}|q_i, \alpha_i)$. First, an auxiliary variable $q_{i+0.5}$ is constructed, which is a beta-distributed random variable with its mode being q_i and the sum of two parameters being $\frac{1}{\alpha_{i+1}}$:

$$v_{i+1} = \frac{1}{\alpha_{i+1}} - 2$$

$$q_{i+0.5} \sim Beta(q_i v_{i+1} + 1, v_{i+1} - q_i v_{i+1} + 1)$$

The conflict probability transition function is then constructed as

$$q_{i+1} \sim q_{i+0.5} + \alpha_{i+1}(o_i - q_{i+0.5})$$

The transition function adopts a classical update rule used in reinforcement learning models, and the learning rate controls the balance between past ($q_{i+0.5}$) and current information (o_i). For the MSIT model, we take the product of the transition functions computed separately for Simon and Flanker predicted conflict. Finally, we consider the likelihood function. Since the trial sequences were designed and re-used for different subjects, the estimated conflict probability would be the same across subjects for the same sequence, but this is inconsistent with the fact that such individual estimates should be subjective and different between participants. We thus incorporated RTs from each subject, which are assumed to be generated through a drift-diffusion process, to estimate a *subjective* conflict probability based on the assumption that a subjects' RT reflects the extent to which they engaged control. RT is assumed to be generated by a diffusion process. We used an abstract version of the drift-diffusion model where the two bounds represent the correct and wrong choice (and not the actual choices). The diffusion process accumulated the *difference* in the evidence between the target and distractor response (Fig. 1c, right), which is smaller for conflict trials and thus leads to longer RTs. We refer to the drift-diffusion likelihood function for RT as $p_{DDM}(v_{si}, v_{fl}, v_{sf}, v_{non-conflict}, vz_{si}, vz_{fl}, a, q_{si}, q_{fl})$ for the MSIT model, and $p_{DDM}(v_{conflict}, v_{stroop\ non-conflict}, vz, a, q_s)$ for the Stroop model (Navarro and Fuss, 2009). The hyperparameters specifying the DDM are boundary separation (a), base drift rates for Simon-only, Flanker-only, both Simon and Flanker present, and non-conflict trials in MSIT ($v_{si}, v_{fl}, v_{sf}, v_{non-conflict}$), base drift rates for conflict and non-conflict trials in Stroop ($v_{conflict}, v_{stroop\ non-conflict}$), and drift rate bias coefficients that scales the conflict probability of Simon, Flanker and Stroop (vz_{si}, vz_{fl}, vz). The effective drift rate is then the sum between the base drift rate and the drift rate bias (see Fig. 1c). Here we made the assumption that conflict prior affects RT by biasing *drift rates* based on a previous work investigating the effect of choice history on RT (Urai et al., 2019). We also assumed that the drift rate diffusion started at the half point of the boundary separation (i.e., $z = 0.5$). With the Markovian assumption, the updating process for the MSIT model is thus given by

$$p(k, \alpha_{i+1}, q_{si,i+1}, q_{fl,i+1} | o_{si,\le i+1}, o_{fl,\le i+1}, RT_{\le i+1}) \propto p(o_{si,i+1}, o_{fl,i+1}, RT_{i+1} | q_{si,i+1}, q_{fl,i+1})$$
$$\iint \left[\int p(k, \alpha_i, q_{si,i}, q_{fl,i} | o_{si,\le i}, o_{fl,\le i}, RT_{\le i}) p(\alpha_{i+1}|\alpha_i, k) d\alpha_i \right] p(q_{si,i+1}|q_{si,i}, \alpha_i) p(q_{fl,i+1}|q_{fl,i}, \alpha_i) dq_{si,i} dq_{fl,i}$$

The updating process for the Stroop model is given by

$$p(k, \alpha_{i+1}, q_{s,i+1} | o_{s,\le i+1}, RT_{\le i+1})$$
$$\propto p(o_{s,i+1}, RT_{i+1} | q_{s,i+1}) \int \left[\int p(k, \alpha_i, q_i | o_{s,\le i}, RT_{\le i}) p(\alpha_{i+1}|\alpha_i, k) d\alpha_i \right] p(q_{s,i+1}|q_{s,i}, \alpha_i) dq_{s,i}$$

The likelihood function is the product of Bernoulli likelihood for trial congruency and DDM likelihood for RT. For MSIT, the likelihood function is given as:

$$p(o_{si,i+1}, o_{fl,i+1}, RT_{i+1} | q_{si,i+1}, q_{fl,i+1}) = (1 - |o_{si,i+1} - q_{si,i+1}|)(1 - |o_{fl,i+1} - q_{fl,i+1}|) p_{DDM}$$

For Stroop, the likelihood function is given as:

$$p(o_{s,i+1}, RT_{i+1} | q_{s,i+1}) = (1 - |o_{s,i+1} - q_{s,i+1}|) p_{DDM}$$

These hyperparameters are estimated using an expectation-maximization (EM) algorithm as shown in earlier work (Jiang et al., 2015). Briefly, the model parameters were at first estimated *without* incorporating the DDM likelihood for RT ("E" step). Hyperparameters were then fit by maximizing the DDM likelihood for the observed RT using the conflict prior (s) obtained ("M" step). The DDM likelihood function with the fitted hyperparameters were then incorporated into the Bayesian updating process ("E" step) to generate a new set of conflict prior (s), which were then used to maximize the DDM likelihood over observed RT again. These steps were repeated until the convergence of both model parameters and hyperparameters (Euclidean distance between parameter vectors from successive iterations < 10^{-5}).

We considered three alternative classes of models: 1) reinforcement learning (RL) models; 2) constant model; 3) Bayesian learning model without RT tuning. For the RL model, we constructed a value function corresponding to the estimated conflict probability, and this estimate is also updated trial-by-trial using a Rescorla-Wagner rule. For MSIT, the update rule is:

$$q_{si,i+1} = q_{si,i} + \alpha\left(o_{si,i+1} - q_{si,i}\right)$$

$$q_{fl,i+1} = q_{fl,i} + \alpha\left(o_{fl,i+1} - q_{fl,i}\right)$$

For Stroop, the update rule is:

$$q_{s,i+1} = q_{s,i} + \alpha\left(o_{s,i+1} - q_{s,i}\right)$$

The free parameter α in the RL models was fit by maximizing the data likelihood (Bernoulli) for trial congruency. For the constant model, q_s, q_{si}, q_{fl} were fit directly by maximizing the data likelihood for trial congruency. For Bayesian conflict learning models without RT tuning, q_s, q_{si}, q_{fl} were estimated online but the likelihood function for RT was not incorporated in the process.

We used the Bayesian Information Criterion (BIC) to compare the RT-tuned Bayesian conflict learning models with the RL models, constant models and the non-RT tuned Bayesian conflict learning models. We compared these models separately for their ability to explain trial congruency and RT. For this analysis, we pooled all data from all sessions and computed the BIC for each model and for each data type (RT or trial congruency), consistent with a previous study (Behrens et al., 2007). Results of model comparisons can be found in Tables S2 and S3.

Selection of neurons

We defined the epochs of interests according to events in the tasks (see Figure 3a for an illustration). The baseline epoch starts at 1.5s before stimulus onset and ends at stimulus onset. This epoch is used to analyze encoding of conflict prior. The ex-ante epoch is anchored to the midpoint of a period of time starting at 100ms after stimulus onset (to account for the minimal delay needed for visual information to reach the MFC) and ending at the time of button presses. We then defined the ex-ante epoch as a 500ms window centered on the midpoint of this period. The rationale for analyzing conflict signals in this epoch is as follows: at the early stage of stimulus processing, information about the different response options is not yet fully processed and hence minimal conflict; the conflict signal should reach its maximum when the different stimulus dimensions that drive competing responses are fully available; and finally, it should subside after a response is selected.

We counted the number of spikes in these epochs and regressed the spike counts against the different regressors (error, conflict, conflict surprise, conflict prior and conflict posterior) using linear regression. For each regressor, we extracted a p value computed from the F test. A neuron was deemed selective for this regressor when $p < 0.05$. For MSIT, since there were both Simon and Flanker conflicts, neurons were selected when regressors related to either Simon or Flanker conflict were significant (e.g., ex-ante conflict cells in MSIT were the union of neurons selective for Simon conflict and Flanker conflict during the ex-ante epoch). To assess whether a neuronal class is significantly present in the population, we derive a null distribution by permuting the relation between spike counts and the regressor of interest for 1000 times. A p value is computed by comparing the true proportion of selected neurons against this null distribution. The 95th interval of the null distribution for each neuronal class is plotted as dotted lines in Figure 3b.

To statistically compare the extent of multiplexing between two groups of cells active in different epochs (Figure 3d), we used the chi-squared test and reported the p-value and effect size of the test.

Single-trial spike train latency

We estimated the onset latency in individual trials using Poisson spike-train analysis (Figure 3e) (Hanes et al., 1995). This method detects the moments when the observed inter-spike intervals (ISI) deviate significantly from that assumed by a constant-rate baseline Poisson process. We used the spike rate averaged across the whole block of experiment as a baseline spike rates for each neuron. This baseline rate was then used to compute a Poisson surprise metric across the spike train. We started our detection algorithm from the onset of stimulus for each trial. For the ex-ante conflict neurons (two columns on the left), we restricted the range in which the detection algorithm looks for bursts to after stimulus onset and before button presses. This is because, by their definition, ex-ante conflict neurons carried a conflict signal before action. For the ex-post conflict neurons (two columns on the right), we restricted the range to 200ms before button presses and before end of trial. We then extract the latency of the first significant burst. 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. For these analyses, we only used the conflict trials as we focused on the single-trial conflict response of selected conflict-encoding neurons.

Correct-related potential (CRP) analyses

We simultaneously also recorded the intracranial electroencephalography (iEEG) while we recorded single unit activity. iEEG data were acquired from low-impedance macro contacts closest to the microwires. We focused on the contacts that were directly placed in dACC and pre-SMA (as confirmed by post-operative imaging, see (Fu et al., 2019)). To extract the CRP, we downsampled the iEEG data to a sampling frequency of 100Hz (using MATLAB “resample”) and then bandpass filtered (0.1Hz-10Hz) the data with a finite impulse response filter (MATLAB function “fir1”). Filtered data were then shifted in time to account for average filter delay (computed using MATLAB function "grpdelay"). We then computed the CRP amplitude for each trial by averaging the filtered iEEG data within [0,250ms] after button presses.

To analyze whether the CRP amplitude was related to conflict and/or RT, we used a linear mixed-effect model, pooling experimental sessions and electrodes. The model in Wilkinson's notation is given by

$$\mathrm{CRP} \sim Stroop\ prior * Stroop\ conflict + RT + (1 + Stroop\ prior + RT | sessionID: subjectID)$$

The fixed effect of $Stroop\ conflict$ is a dummy variable indicating whether a trial is a conflict trial (value = 1) or not (value = 0). We analyzed the relation between spike counts of prior neurons in dACC and pre-SMA and the simultaneously recorded CRP using a Poisson mixed-effect model. Spike counts were gathered using a 500ms bin swept across the trial in steps of 25ms. The model was computed for spike counts in each bin. The full model is given by:

$$\text{Spike counts}(t) \sim CRP + Stroop\ prior + RT + (1|cellID)$$

To determine the statistical significance of each fixed effect, we compare the full model with a reduced model with the fixed effect of interest removed and tested the likelihood ratio between the full and reduced model using a likelihood ratio test. For the CRP-spike count relation model, the reduced model is the same as the full model except that the fixed effect of CRP is removed. We plotted the likelihood ratio from this model comparison as a function of time in Figure 3I. The p-values were obtained from the likelihood ratio tests and corrected using false discovery rate method.

Detrended fluctuation analysis

Detrended fluctuation analysis (DFA) was first developed by Peng and colleagues (Hardstone et al., 2012; Peng et al., 1994) to quantify long-range temporal correlations (LRTC). We use DFA to quantify the extent of LRTC in baseline spike counts on the scale of *trials.* First, the cumulative sum of the spike counts during baseline was computed. To be consistent with prior literature we refer to this cumulative sum as the *signal profile*. A set of trial window sizes were defined between the lower bound of 4 trials and the upper bound of the block length. For each window size, we then partitioned the signal profile into a series of data snippets. Partitioning was done such that two adjacent snippets had an overlap of half the window size. We then removed the linear trend from each data snipped (using least square regression) and computed the standard deviation across time. The mean of the standard deviations across all snippets of identical window size was then computed (y axis of Fig. 4f). Finally, the mean standard deviations were regressed linearly against the logarithmically scaled time windows and the slope was extracted as the DFA α value (Fig. 4f shows the fluctuations as a function of logarithmically transformed trial window sizes for two example neurons).
For Fig 4a-b, we tested the relation between a neuron's baseline DFA α value and its tendency to encode conflict prior. To avoid selection bias, we split trials into two sets of equal size, with one

half consisting of a consecutive run of trials. This is because DFA is used for time series data and thus required the data be consecutive and temporally ordered. For randomization, we first randomly sampled one trial from the first half of the block. Then a consecutive run of trials starting with this randomly picked trial as the starting point were extracted. The consecutive set was used to compute DFA α value while the rest of the trials were used to correlate with conflict prior (Simon, Flanker or Stroop) using Spearman rank correlation.

Decoding analysis (Support-vector machine)

Data were aggregated from different experimental sessions to create pseudo-population data matrices. We constructed for each trial a peri-stimulus time histogram (PSTH) using 500ms bins in steps of 25ms. For all conflict or conflict prior -related decoding, we used correct trials only. Since different behavioral sessions had different number of trials (some subjects participated in less sessions than the others), we subsampled the same number of trials from each condition for each neuron and repeat this process 50 times. For error decoding, we subsampled 10 error trials and 10 correct conflict trials for Stroop and 10 correct sf trials for MSIT. Since most errors occurred on high conflict trials, these contrasts isolate the effect of error while controlling for the effect of conflict. For conflict decoding, we subsampled 30 trials from each conflict conditions: conflict and non-conflict trials for Stroop; Simon and non-Simon trials, Flanker and non-Flanker trials for MSIT. For each time bin, we performed 5-fold cross validation using LIBSVM (Chang and Lin, 2011). We used the linear kernel and set the c parameter to 1 for all analyses. In brief, trials were first randomly split into 5 equal parts; each part was used in turn as the testing data while the rest of the four parts were used as training data. Decoding accuracy was the proportion of correct classifications among the 250 samples (50 resamples x 5 folds). Note that the resampling was done once to generate testing and training sets for the whole time series and used for both within-time and across-time decoding. For within-time decoding (All plots with dotted lines in Fig. 5), the SVM classifier trained using the training data from each time bin was tested using the testing data from the same time bin. For cross-temporal decoding (temporal generalization; all plots with solid lines in Fig. 5), the SVM classifier trained using the training data from each time bin were tested across the trial using testing data gathered from other time bins.

Reaction times equalization

For analyses shown in Figures S6c and S7a-d, we selected a subset of trials from each condition so that the RTs did not differ significantly across conditions (e.g., equalizing RTs between conflict and non-conflict trials in the Stroop data). Here we detail the RT equalization procedure we used to create "RT equalized sets". We first selected a condition as the "anchor" condition. We sorted the RTs of the anchor condition, and for each RT we searched in the target (to-be-equalized) condition(s) for a trial whose RT did not differ from the anchor RT by more than 0.1s. If all RTs in the target condition differed from the anchor RT by > 0.1s, the anchor RT was not included in the RT equalized set. Once selected, the anchor RT and the target RT were both removed from the original set to ensure that no trials were included twice in any RT equalized trial sets. This procedure was repeated until one of the conditions considered were emptied. We confirmed post hoc that RT equalization was successful by testing whether RTs were not significantly different using ANOVA ($p > 0.5$ for all the RT equalized sets).

Decoding analysis (population activity vectors and demixed PCA)

Data were aggregated from different experimental sessions to create a pseudo-population. We randomly selected one trial for each neuron from one condition and concatenate the data from each neuron to form a single-trial testing data matrix. The rest of the trials were averaged for each condition and concatenated to form a training data matrix. Coding dimensions were defined based on the condition-averaged training data. To define the coding dimensions used to decode conflict conditions within MSIT, we used the population activity vectors (a high dimensional vector in the raw firing rate space) defined by the difference between the two condition means. To define coding dimensions for the cross-task decoding problems we used dPCA to extract demixed principal components (dPC). Details of which trials were used to define the coding dimensions used to generate Figures 6,7, S6 and S7 are given in the sections to follow. Both testing and training data were projected onto the identified coding dimensions. The labels for testing data were assigned according to the label of the nearest neighbor of the training data. To test condition generalization, we projected the testing data from one pair of conditions to a coding dimension defined by another pair of conditions (e.g., a Simon trial and non-Simon trial projected to the population vector flanked by Flanker and non-Flanker trial averages) and classified using the labels of the nearest projected training data. This decoding procedure was repeated 1000 times (resulting in 1000 single-trial testing data matrices and the corresponding training data matrices), and the decoding accuracy was defined by the proportion of correct classifications among these 1000 repetitions. To determine statistical significance, we permuted the trial labels for 500 times and for each permutation, we repeated all above steps to generate a null distribution. A p-value was computed from comparing the true decoding accuracy with the null distribution.

Pseudo-population matrices for MSIT analyses

For Figure 6a, we formed the pseudo-population data matrix by taking the average of spike counts within the ex-ante or ex-post (1s after button presses) epoch across all Simon-only, Flanker-only, Simon+ Flanker, and non-conflict trials, respectively. We then used PCA on this condition-averaged data matrix to extract the three principal components (PC) that explained most variance to visualize the geometric arrangement of the four conflict types. For Figure 6f-h, trials were binned by quartiles of prior and posterior into four bins separately. However, because conflict prior was updated into conflict posterior after each button press, binning priors does not guarantee that the posteriors would fall into the same bins. This is because updating is specific to each behavioral session and thus differs between neurons. Averaging trials using only bins formed by prior quartiles would thus mix trials with different levels of posterior for each neuron. To avoid this problem, we thus formed the data matrix (which now includes the time dimension rather than a single ROI; spikes were counted in 500ms bins swept across the whole trial in steps of 25ms; spike trains were aligned to button presses) by concatenating two submatrices: one that was constructed by averaging trials within bins defined by prior quartiles using data *before* button presses, and one that was constructed from averaging trials within bins defined by posterior quartiles for neural data *after* the button press. We then used PCA to find the three PCs that explained the most variance for this matrix. The concatenated data matrix was then projected onto these PCs to generate the visualization of trajectory corresponding to prior/posterior levels.

Vectors in the state space to quantify population geometry within MSIT

We next describe how coding dimensions were defined in each case using population activity vectors in the raw firing rate space. For Figure 6b, the coding dimension was the population vector flanked by the trial averages of sf and non-conflict trials (Fig. 6a, dashed lines). Classifications were carried out between pairs of conflict conditions (e.g., between si and fl trials) as detailed above. For Figure 6c, we took a bin-wise approach to investigate whether Simon conflict representation generalize to Flanker representation, and vice versa. For this, we split trials into four non-overlapping groups: Simon, Flanker, non-Simon, non-Flanker trial sets. We split sf and non-conflict trials randomly in half. One half of sf trials were pooled with si trials to form the Simon trial set, and one half of non-conflict trials were pooled with fl trials to form the non-Simon trial set. The other half of sf trials were then pooled with fl trials to form the Flanker trial set, and the other half of non-conflict trials were pooled with si trials to form non-Flanker trial set. Using these trial sets, for each time bin we extracted two coding dimensions from the training data: one population vector flanked by trial averages of Simon and non-Simon trials (Simon coding dimension), and one population vector flanked by the trial averages of Flanker and non-Flanker trials (Flanker coding dimension). We then projected the testing data from Simon/non-Simon trials onto the Flanker coding dimension and classified the testing data using the closest projected training data, and vice versa. For details of this classification procedure see above paragraph. This assesses the extent to which coding of Simon and Flanker conflict is abstract.

Compositionality of conflict representation

For Figure 6d, the coding dimension were taken to be the blue and orange edges as shown in Figure 6a. The purpose of this analysis is to assess to what extent the representation of conflict is compositional (within a task). We assumed that in the neuronal firing rate space, the representation of Simon/Flanker conflict is a vector pointing from non-conflict trial averages to the si/fl trial averages. Compositionality of such conflict representation would imply that the sf representation (vector pointing from non-conflict trial average to the sf trial average) is equal to the sum of the Simon and Flanker representations. According to the parallelogram law of vector addition, this then corresponds to the blue and orange edges in Figure 6a forming a parallelogram. We tested the extent of parallelism in the data using decoding. The coding dimensions here were defined by the following population vectors using training data: one flanked by non-conflict and si trial averages (Fig. 6a, blue), one flanked by fl and sf trial averages (blue), one flanked by fl and non-conflict trial averages (orange) and one flanked by si and sf trial averages (orange). Left-out testing data from conditions flanking one of the blue or orange pair of edges were then projected to the other edge in the pair and classified by the training data defining this edge. For example, single-trial testing data of non-conflict and si trials were projected to the coding dimension flanked by fl and sf trial averages and were classified by fl or sf trial averages.

Relationship of single neuron tuning with parallelism in geometry

For Figure 6e and Figure S6e, the goal is to investigate the relation between the nonlinearity in single neuron conflict coding and the deviation from perfect compositionality in state space representation of conflict. We denote the state-space representation of Simon and Flanker -only conflict as the population vectors flanked by the trial averages of si and non-conflict and by the trial averages of fl and non-conflict. We refer to the state space location occupied by the linear sum of Simon and Flanker representation defined above as "s+f". The deviation from perfect compositionality is then given by the population vector flanked by "sf" and "s+f". The loading of "sf" to the "s+f" vector reflects the single neuron contribution to the deviation at the population level. To quantify nonlinearity of conflict coding for each neuron, we first regressed the spike counts in the ex-ante or ex-post epoch (1s) against three fixed effects: a Simon effect (dummy variable indicating the presence or absence of Simon conflict on a trial), a Flanker effect (dummy variable indicating the presence or absence of Flanker conflict on a trial) and the interaction term between these two. We extracted the F statistic related to the interaction term, which captures the effect of nonlinear mixing of Simon and Flanker conflict. We then extracted a population vector flanked by the sf trial average and and "s+f", the sum of two population vectors one flanked by trial averages of si trials and non-conflict trials, and one flanked by trial averages of fl trials and non-conflict trials. We then correlated the loading of "sf" - "s+f" vector and the F statistics from a particular neuron.

Quantification of state space dynamics

For Figure 6i, we binned spike counts using 250ms bins swept across the trial in steps of 10ms. The state-space speed was defined to be the Euclidean distance between population vectors of adjacent time bins divided by the step size. We averaged the state-space speed across time within an epoch. We also computed the Euclidean distance between pairs of trajectories (1^{st} and 2^{nd}, 2^{nd} and 3^{rd},3^{rd} and 4^{th}) and averaged this across trajectories and across time bins within an epoch. State-space speed and the averaged distance between trajectories were plotted against each other in Figure 6I. Our method for extracting speed in state-space follows prior work (Stokes et al., 2013).

Testing ordinal relationship of prior/posterior projections

We analyzed the ordinal relation between neural projections of prior/posterior as shown in Figure 6j-l. PCA axes encoding prior/posterior variance were extracted from spike count data collected in ROIs (baseline for prior and the ex-post epoch (0-1s after button presses) for posterior). Since prior/posterior is continuously valued, we created four trial conditions by binning the trials using quartiles of prior/posterior. For each type of prior or posterior (Simon, Flanker and Stroop), we projected the left-out trial (not used for computing the PCA axis) onto the PCA axis for each trial condition and this procedure was repeated 1000 times, yielding 1000 projected values for each trial condition. We then regressed the projected values (concatenated into a vector) against their trial condition labels (1st,2nd,3rd,4th quartile bins) using a multinomial logistic regression with the assumption of ordinal relation between trial groups. Essentially, we were testing whether the out-of-sample project values can reliably predict the trial condition they belong to assuming that the conditions were ordinal. We reported the p-value and t-statistic of the effect of projected values.

Demixed Principal Component Analyses (dPCA)

We used dPCA to extract task-invariant representation of performance monitoring signals. For Figures 7 and S7, we investigated task-invariant coding of error, conflict and conflict prior separately, resulting in three separate optimization problems. For Figure S6, we investigated the invariance coding of error and conflict. Analyses on conflict and conflict prior used only correct trials. We used dPCA as described previously (Kobak et al., 2016), with the following adaptions made for our purposes. The dPCA algorithm first decomposes population neural activity into marginalized data matrices with respect to the variables of interest. For analyses in Figure 7a-c, we constructed the marginalized population activity (referred generically as $\overline{X_{\phi}}$) with respect to error ($\overline{X_{error}}$ in Fig. 7a) or conflict ($\overline{X_{conflict}}$, Fig. 7b-c) by marginalizing out time and task dimensions (denoted by "$\langle\cdot\rangle_{task,t}$").

For interpretability, we investigated whether the neural representation is abstract across tasks separately between Stroop and Simon conflict ("$s\ \&\ si\ conflict$" is the task-invariant dimension indicating presence of absence of conflict for both tasks) and between Stroop and Flanker ("$s\ \&\ fl\ conflict$" is the task-invariant dimension indicating presence of absence of conflict for both tasks). Set up this way, the "task" dimension captures variance related to task set differences (Stroop vs. MSIT). To compute marginalized averages, we use N-dimensional population activity

$$\overline{X_{Stroop\ error\ \&\ MSIT\ error}} = \langle r(error, task, t) - \bar{r}(t)\rangle_{task,t}$$

$$\overline{X_{s\ \&\ si\ conflict}} = \langle r(s\ \&\ si\ conflict, task, t) - \bar{r}(t)\rangle_{task,t}$$

$$\overline{X_{s\ \&\ fl\ conflict}} = \langle r(s\ \&\ fl\ conflict, task, t) - \bar{r}(t)\rangle_{task,t}$$

, where $\bar{r}(t)$ is the firing rate averaged across trials and time bins. For Figure S7a-c, these definitions are the same except that RT equalized trial sets were used.

For Figure S6a-b, we sought a common coding dimension between error and conflict separately for MSIT and Stroop, by marginalizing out the information about time and which pair of conditions were contrasted ("contrast" indicator, for MSIT it indicates whether the contrast considered is sf vs. non-conflict or error sf vs. correct sf; for Stroop it indicates whether the contrast considered is correct conflict vs. correct non-conflict or error vs. correct conflict).

we constructed the marginalized population activity with respect to error vs conflict in both MSIT ($\overline{X_{MSIT\ error/sf}}$ in Fig. S6a, $MSIT\ error/sf$ is the contrast-invariant dimension indicating presence or absence of errors and presence or absence of sf conflict) and Stroop ($\overline{X_{Stroop\ error/conflict}}$ in Fig. S6b, $Stroop\ error/conflict$ is the contrast-invariant dimension indicating presence or absence of errors and presence or absence of sf conflict) as follows:

$$\overline{X_{MSIT\ error/sf}} = \langle r(MSIT\ error/sf, contrast, t) - \bar{r}(t)\rangle_{contrast,t}$$

$$\overline{X_{Stroop\ error/conflict}} = \langle r(Stroop\ error/conflict, contrast, t) - \bar{r}(t)\rangle_{contrast,t}$$

For analyses in Figure 7d-f and Figure S7e-f investigating task-invariant coding of conflict prior, we used data from a single ROI (ex-ante or ex-epoch) and hence only the task but not time dimension was marginalized out. Here again for interpretability, we investigated cross-task representation between Stroop prior and Simon prior and between Stroop prior and Flanker prior separately, ensuring that the task dimension captures task set difference

$$\overline{X_{s\ \&\ si\ conflict}} = \langle r(s\ \&\ si\ conflict, task) - \bar{r}(t)\rangle_{task}$$

$$\overline{X_{s\ \&\ fl\ conflict}} = \langle r(s\ \&\ fl\ conflict, task) - \bar{r}(t)\rangle_{task}$$

$$\overline{X_{s\ \&\ si\ prior}} = \langle r(s\ \&\ si\ conflict, task) - \bar{r}(t)\rangle_{task}$$

$$\overline{X_{s\ \&\ fl\ prior}} = \langle r(s\ \&\ fl\ conflict, task) - \bar{r}(t)\rangle_{task}$$

For analyses in Figure S7d, these definitions are the same except that RT equalized trial sets were used.

The algorithm then finds encoding (F_ϕ) and decoding (D_ϕ) matrices separately for each marginalized averages using the regularized reduced-rank regression:

$$L_\phi = \left\|\overline{X_\phi} - F_\phi D_\phi \bar{X}\right\|^2 + \mu\left\|F_\phi D_\phi\right\|^2$$

We assigned a regularization coefficient μ to avoid overfitting ($\mu = 6e^{-6}$ determined from results reported in (Kobak et al., 2016)). We used the columns of D_ϕ as the demixed principal components (dPC) and projected N-dimensional data (single-trial data for testing and trial-averaged data for training) to these dPCs. The numerical values of D_ϕ reflects the contribution for each neuron to task-invariant representation.

To test the statistical significance of coding dimensions, we randomly chose one trial for each trial type (e.g., one error trial and one correct trial) and constructed a single-trial activity matrix X_{test}. We then used the remaining trials to form the trial-averaged training data $\overline{X_{train}}$, which is used to find the dPCA coding dimensions. The left-out single-trial data X_{test} is then projected onto the first coding dimension that captures the most variance computed from X_{train}, and classified according to the closest class mean. We repeated this procedure 1000 times and determined the decoding accuracy as the proportion of correct classification among the 1000 test trials. We then generated the null distribution by shuffling the trial labels and then repeated the decoding procedure 500 times. For Figure 7d-f, statistical significance is determined by comparing the true decoding accuracy with this null distribution. For Figure 7a-c, statistical significance is determined by the cluster-based permutation test using this null distribution (Maris and Oostenveld, 2007). The fraction of explained variance (Bars in Figures 7a-c,e-f and S7a-c,e-f) for each marginalization is given by:

$$R_{\phi}{}^{2} = \frac{\left\|\overline{X_{\phi}}\right\|^{2} - \left\|\overline{X_{\phi}} - F_{\phi} D_{\phi} \bar{X}\right\|^{2}}{\left\|\overline{X_{\phi}}\right\|^{2}}$$

For analyses in Figure 7g-j and Figure S7 g-k, we first quantify for each single neuron its task-invariant coding strength of error, conflict or conflict prior within a certain ROI. Spike counts within the baseline, ex-ante or ex-post epochs from MSIT and Stroop were concatenated and were regressed against three fixed effects: a cognitive effect (trial outcome, trial congruency or conflict prior), a task effect (a dummy variable with value 1 for MSIT and value 0 for Stroop) and an interaction between these two. The signed effect size is taken to be the t-statistic computed from this linear regression. The t-statistic related to the cognitive effect characterizes the strength of task-invariant coding of cognitive variables (error, conflict and conflict prior). The task dependency of such coding is captured by the t-statistic of the interaction term (which indicates that a neuron exhibits non-linear mixing).

Neurons with a significant cognitive effect but a non-significant task effect and a non-significant interaction term were classified as "task-invariant" neurons. Neurons with a significant interaction term are classified as "task-dependent" neurons.

References

Aarts, E., Verhage, M., Veenvliet, J.V., Dolan, C.V., and van der Sluis, S. (2014). A solution to dependency: using multilevel analysis to accommodate nested data. Nat. Neurosci. *17*, 491–496.

Aron, A.R., Behrens, T.E., Smith, S., Frank, M.J., and Poldrack, R.A. (2007). Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI. J. Neurosci. Off. J. Soc. Neurosci. *27*, 3743–3752.

Bach, D.R., and Dayan, P. (2017). Algorithms for survival: a comparative perspective on emotions. Nat. Rev. Neurosci. *18*, 311–319.

Bean, B.P. (2007). The action potential in mammalian central neurons. Nat. Rev. Neurosci. *8*, 451–465.

Behrens, T.E.J., Woolrich, M.W., Walton, M.E., and Rushworth, M.F.S. (2007). Learning the value of information in an uncertain world. Nat. Neurosci. *10*, 1214–1221.

Bernacchia, A., Seo, H., Lee, D., and Wang, X.-J. (2011). A reservoir of time constants for memory traces in cortical neurons. Nat. Neurosci. *14*, 366–372.

Bernardi, S., Benna, M.K., Rigotti, M., Munuera, J., Fusi, S., and Salzman, C.D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. Cell *183*, 954-967.e21.

Bonini, F., Burle, B., Liégeois-Chauvel, C., Régis, J., Chauvel, P., and Vidal, F. (2014). Action Monitoring and Medial Frontal Cortex: Leading Role of Supplementary Motor Area. Science *343*, 888–891.

Botvinick, M.M., Braver, T.S., Barch, D.M., Carter, C.S., and Cohen, J.D. (2001). Conflict monitoring and cognitive control. Psychol. Rev. *108*, 624–652.

Brainard, D.H. (1997). The Psychophysics Toolbox. Spat. Vis. *10*, 433–436.

Braver, T.S. (2012). The variable nature of cognitive control: a dual mechanisms framework. Trends Cogn. Sci. *16*, 106–113.

Bush, G., and 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*, 308–313.

Buzsáki, G., Anastassiou, C.A., and Koch, C. (2012). The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. Nat. Rev. Neurosci. *13*, 407–420.

Carter, C.S., Braver, T.S., Barch, D.M., Botvinick, M.M., Noll, D., and Cohen, J.D. (1998). Anterior Cingulate Cortex, Error Detection, and the Online Monitoring of Performance. Science *280*, 747–749.

Carter, C.S., Macdonald, A.M., Botvinick, M., Ross, L.L., Stenger, V.A., Noll, D., and Cohen, J.D. (2000). Parsing executive processes: Strategic vs. evaluative functions of the anterior cingulate cortex. Proc. Natl. Acad. Sci. *97*, 1944–1948.

Cavanagh, J.F., Wiecki, T.V., Cohen, M.X., Figueroa, C.M., Samanta, J., Sherman, S.J., and Frank, M.J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. Nat. Neurosci. *14*, 1462–1467.

Cavanagh, S.E., Wallis, J.D., Kennerley, S.W., and Hunt, L.T. (2016). Autocorrelation structure at rest predicts value correlates of single neurons during reward-guided choice. ELife *5*, e18937.

Cavanagh, S.E., Towers, J.P., Wallis, J.D., Hunt, L.T., and Kennerley, S.W. (2018). Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. Nat. Commun. *9*, 3498.

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. *2*, 27:1-27:27.

Crone, E.A., Somsen, R.J.M., Beek, B.V., and Molen, M.W.V.D. (2004). Heart rate and skin conductance analysis of antecendents and consequences of decision making. Psychophysiology *41*, 531–540.

Crosson, B., Sadek, J.R., Bobholz, J.A., Gökçay, D., Mohr, C.M., Leonard, C.M., Maron, L., Auerbach, E.J., Browd, S.R., Freeman, A.J., et al. (1999). Activity in the Paracingulate and Cingulate Sulci during Word Generation: An fMRI Study of Functional Anatomy. Cereb. Cortex *9*, 307–316.

Danielmeier, C., Eichele, T., Forstmann, B.U., Tittgemeyer, M., and Ullsperger, M. (2011). Posterior Medial Frontal Cortex Activity Predicts Post-Error Adaptations in Task-Related Visual and Motor Areas. J. Neurosci. *31*, 1780–1789.

Darlington, T.R., Beck, J.M., and Lisberger, S.G. (2018). Neural implementation of Bayesian inference in a sensorimotor behavior. Nat. Neurosci. *21*, 1442–1451.

DiCarlo, J.J., and Cox, D.D. (2007). Untangling invariant object recognition. Trends Cogn. Sci. *11*, 333–341.

Dubois, J., de Berker, A.O., and Tsao, D.Y. (2015). Single-unit recordings in the macaque face patch system reveal limitations of fMRI MVPA. J. Neurosci. Off. J. Soc. Neurosci. *35*, 2791–2802.

Duthoo, W., Abrahamse, E.L., Braem, S., Boehler, C.N., and Notebaert, W. (2014). The heterogeneous world of congruency sequence effects: an update. Front. Psychol. *5*.

Ebitz, R.B., and Platt, M.L. (2015). Neuronal Activity in Primate Dorsal Anterior Cingulate Cortex Signals Task Conflict and Predicts Adjustments in Pupil-Linked Arousal. Neuron *85*, 628–640.

Egner, T. (2007). Congruency sequence effects and cognitive control. Cogn. Affect. Behav. Neurosci. *7*, 380–390.

Egner, T., and Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. Nat. Neurosci. *8*, 1784–1790.

Eldar, E., Rutledge, R.B., Dolan, R.J., and Niv, Y. (2016). Mood as Representation of Momentum. Trends Cogn. Sci. *20*, 15–24.

Fan, J., Flombaum, J.I., McCandliss, B.D., Thomas, K.M., and Posner, M.I. (2003). Cognitive and Brain Consequences of Conflict. NeuroImage *18*, 42–57.

Friston, K.J. (2002). Bayesian Estimation of Dynamical Systems: An Application to fMRI. NeuroImage *16*, 513–530.

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

Fusi, S., Miller, E.K., and Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. Curr. Opin. Neurobiol. *37*, 66–74.

Gratton, G., Coles, M.G., and Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. J. Exp. Psychol. Gen. *121*, 480–506.

Hanes, D.P., Thompson, K.G., and 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.

Hardstone, R., Poil, S.-S., Schiavone, G., Jansen, R., Nikulin, V.V., Mansvelder, H.D., and Linkenkaer-Hansen, K. (2012). Detrended Fluctuation Analysis: A Scale-Free View on Neuronal Oscillations. Front. Physiol. *3*.

Heilbronner, S.R., and Hayden, B.Y. (2016). Dorsal Anterior Cingulate Cortex: A Bottom-Up View. Annu. Rev. Neurosci. *39*, 149–170.

Heitz, R.P., and Schall, J.D. (2012). Neural Mechanisms of Speed-Accuracy Tradeoff. Neuron *76*, 616–628.

Herrera, B., Sajad, A., Woodman, G.F., Schall, J.D., and Riera, J.J. (2020). A Minimal Biophysical Model of Neocortical Pyramidal Cells: Implications for Frontal Cortex Microcircuitry and Field Potential Generation. J. Neurosci. *40*, 8513–8529.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. (2018). Towards a Definition of Disentangled Representations. ArXiv181202230 Cs Stat.

Ide, J.S., Shenoy, P., Yu, A.J., and Li, C.R. (2013). Bayesian Prediction and Evaluation in the Anterior Cingulate Cortex. J. Neurosci. *33*, 2039–2047.

Ito, S., Stuphorn, V., Brown, J.W., and Schall, J.D. (2003). Performance Monitoring by the Anterior Cingulate Cortex During Saccade Countermanding. Science *302*, 120–122.

Jiang, J., and Egner, T. (2014). Using Neural Pattern Classifiers to Quantify the Modularity of Conflict–Control Mechanisms in the Human Brain. Cereb. Cortex *24*, 1793–1805.

Jiang, J., Heller, K., and Egner, T. (2014). Bayesian modeling of flexible cognitive control. Neurosci. Biobehav. Rev. *46*, 30–43.

Jiang, J., Beck, J., Heller, K., and Egner, T. (2015). An insula-frontostriatal network mediates flexible cognitive control by adaptively predicting changing control demands. Nat. Commun. *6*, 8165.

Kawai, T., Yamada, H., Sato, N., Takada, M., and Matsumoto, M. (2019). Preferential Representation of Past Outcome Information and Future Choice Behavior by Putative Inhibitory

Interneurons Rather Than Putative Pyramidal Neurons in the Primate Dorsal Anterior Cingulate Cortex. Cereb. Cortex *29*, 2339–2352.

Kerns, J.G., Cohen, J.D., MacDonald, A.W., Cho, R.Y., Stenger, V.A., and Carter, C.S. (2004). Anterior Cingulate Conflict Monitoring and Adjustments in Control. Science *303*, 1023–1026.

King, J.A., Korb, F.M., von Cramon, D.Y., and Ullsperger, M. (2010). Post-Error Behavioral Adjustments Are Facilitated by Activation and Suppression of Task-Relevant and Task-Irrelevant Information Processing. J. Neurosci. *30*, 12759–12769.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C.E., Kepecs, A., Mainen, Z.F., Qi, X.-L., Romo, R., Uchida, N., and Machens, C.K. (2016). Demixed principal component analysis of neural population data. ELife *5*, e10989.

de Leeuw, J.R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. Behav. Res. Methods *47*, 1–12.

Liu, X., Banich, M.T., Jacobson, B.L., and Tanabe, J.L. (2004). Common and distinct neural substrates of attentional control in an integrated Simon and spatial Stroop task as assessed by event-related fMRI. NeuroImage *22*, 1097–1106.

Lo, C.-C., and Wang, X.-J. (2006). Cortico–basal ganglia circuit mechanism for a decision threshold in reaction time tasks. Nat. Neurosci. *9*, 956–963.

Logan, G.D., and Zbrodoff, N.J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. Mem. Cognit. *7*, 166–174.

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods *164*, 177–190.

McDougle, S.D., Boggess, M.J., Crossley, M.J., Parvin, D., Ivry, R.B., and Taylor, J.A. (2016). Credit assignment in movement-dependent reinforcement learning. Proc. Natl. Acad. Sci. *113*, 6797–6802.

Miller, E.K., and Cohen, J.D. (2001). An Integrative Theory of Prefrontal Cortex Function. Annu. Rev. Neurosci. *24*, 167–202.

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

Minxha, J., Adolphs, R., Fusi, S., Mamelak, A.N., and Rutishauser, U. (2020). Flexible recruitment of memory-based choice representations by the human medial frontal cortex. Science *368*.

Morales, J., Lau, H., and Fleming, S.M. (2018). Domain-General and Domain-Specific Patterns of Activity Supporting Metacognition in Human Prefrontal Cortex. J. Neurosci. *38*, 3534–3546.

Mosher, C.P., Wei, Y., Kamiński, J., Nandi, A., Mamelak, A.N., Anastassiou, C.A., and Rutishauser, U. (2020). Cellular Classes in the Human Brain Revealed In Vivo by Heartbeat-Related Modulation of the Extracellular Action Potential Waveform. Cell Rep. *30*, 3536-3551.e6.

Murphy, P.R., Boonstra, E., and Nieuwenhuis, S. (2016). Global gain modulation generates time-dependent urgency during perceptual choice in humans. Nat. Commun. *7*, 13526.

Navarro, D.J., and Fuss, I.G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. J. Math. Psychol. *53*, 222–230.

Niv, Y., Daw, N.D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology (Berl.) *191*, 507–520.

Pedersen, M.L., Frank, M.J., and Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. Psychon. Bull. Rev. *24*, 1234–1251.

Peng, C.K., Buldyrev, S.V., Havlin, S., Simons, M., Stanley, H.E., and Goldberger, A.L. (1994). Mosaic organization of DNA nucleotides. Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top. *49*, 1685–1689.

Pouget, P., Logan, G.D., Palmeri, T.J., Boucher, L., Paré, M., and Schall, J.D. (2011). Neural basis of adaptive response time adjustment during saccade countermanding. J. Neurosci. Off. J. Soc. Neurosci. *31*, 12604–12612.

Purcell, B.A., and Kiani, R. (2016). Neural Mechanisms of Post-error Adjustments of Decision Policy in Parietal Cortex. Neuron *89*, 658–671.

Rigotti, M., Barak, O., Warden, M.R., Wang, X.-J., Daw, N.D., Miller, E.K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature *497*, 585–590.

Rutishauser, U., Schuman, E.M., and 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*, 204–224.

Sajad, A., Godlove, D.C., and Schall, J.D. (2019). Cortical microcircuitry of performance monitoring. Nat. Neurosci. *22*, 265–274.

Sarafyazd, M., and Jazayeri, M. (2019). Hierarchical reasoning by neural circuits in the frontal cortex. Science *364*.

Schmidt, J.R., and De Houwer, J. (2011). Now you see it, now you don’t: Controlling for contingencies and stimulus repetitions eliminates the Gratton effect. Acta Psychol. (Amst.) *138*, 176–186.

Shackman, A.J., Salomons, T.V., Slagter, H.A., Fox, A.S., Winter, J.J., and Davidson, R.J. (2011). The integration of negative affect, pain and cognitive control in the cingulate cortex. Nat. Rev. Neurosci. *12*, 154–167.

Shenhav, A., Botvinick, M.M., and Cohen, J.D. (2013). The Expected Value of Control: An Integrative Theory of Anterior Cingulate Cortex Function. Neuron *79*, 217–240.

Sheth, S.A., Mian, M.K., Patel, S.R., Asaad, W.F., Williams, Z.M., Dougherty, D.D., Bush, G., and Eskandar, E.N. (2012). Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. Nature *488*, 218–221.

Sohn, H., Narain, D., Meirhaeghe, N., and Jazayeri, M. (2019). Bayesian Computation through Cortical Latent Dynamics. Neuron *103*, 934-947.e5.

Stokes, M.G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., and Duncan, J. (2013). Dynamic Coding for Cognitive Control in Prefrontal Cortex. Neuron *78*, 364–375.

Stroop, J.R. (1935). Studies of interference in serial verbal reactions. J. Exp. Psychol. *18*, 643–662.

Stuphorn, V., Taylor, T.L., and Schall, J.D. (2000). Performance monitoring by the supplementary eye field. Nature *408*, 857–860.

Tang, H., Yu, H.-Y., Chou, C.-C., Crone, N.E., Madsen, J.R., Anderson, W.S., and Kreiman, G. (2016). Cascade of neural processing orchestrates cognitive control in human frontal cortex. ELife *5*, e12352.

Thura, D., and Cisek, P. (2017). The Basal Ganglia Do Not Select Reach Targets but Control the Urgency of Commitment. Neuron *95*, 1160-1170.e5.

Tzelgov, J., Henik, A., and Berger, J. (1992). Controlling Stroop effects by manipulating expectations for color words. Mem. Cognit. *20*, 727–735.

Ullsperger, M. (2017). Neural bases of performance monitoring. In The Wiley Handbook of Cognitive Control, (Wiley Blackwell), pp. 292–313.

Ullsperger, M., Danielmeier, C., and Jocham, G. (2014). Neurophysiology of Performance Monitoring and Adaptive Behavior. Physiol. Rev. *94*, 35–79.

Urai, A.E., de Gee, J.W., Tsetsos, K., and Donner, T.H. (2019). Choice history biases subsequent evidence accumulation. ELife *8*, e46331.

Vigneswaran, G., Kraskov, A., and Lemon, R.N. (2011). Large Identified Pyramidal Cells in Macaque Motor and Premotor Cortex Exhibit “Thin Spikes”: Implications for Cell Type Classification. J. Neurosci. *31*, 14235–14242.

Vogt, B.A., Nimchinsky, E.A., Vogt, L.J., and Hof, P.R. (1995). Human cingulate cortex: Surface features, flat maps, and cytoarchitecture. J. Comp. Neurol. *359*, 490–506.

Wang, X.-J. (2002). Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. Neuron *36*, 955–968.

Wang, J., Narain, D., Hosseini, E.A., and Jazayeri, M. (2018a). Flexible timing by temporal scaling of cortical responses. Nat. Neurosci. *21*, 102–110.

Wang, S., Mamelak, A.N., Adolphs, R., and Rutishauser, U. (2018b). Encoding of Target Detection during Visual Search by Single Neurons in the Human Brain. Curr. Biol. *28*, 2058-2069.e4.

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

Wiecki, T.V., Sofer, I., and Frank, M.J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. Front. Neuroinformatics *7*.

Woodman, G.F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. Atten. Percept. Psychophys. *72*, 2031–2046.

Yeung, N., and Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. Philos. Trans. R. Soc. B Biol. Sci. *367*, 1310–1321.

Yu, A.J., and Cohen, J.D. (2008). Sequential effects: Superstition or rational behavior? Adv. Neural Inf. Process. Syst. *21*, 1873–1880.