RESEARCH ARTICLE SUMMARY

NEUROSCIENCE

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

Zhongzheng Fu*, **Danielle Beam, Jeffrey M. Chung, Chrystal M. Reed, Adam N. Mamelak, Ralph Adolphs, Ueli Rutishauser***

INTRODUCTION: Monitoring our own behavior without explicit feedback is a prominent human ability, enabling us to evaluate whether we made an error, are experiencing conflict, or are in an easy or difficult task environment. In real-life situations, we often need to rapidly learn to perform novel tasks with minimal instruction and time for practice, which requires domain-general performance monitoring that functions "out of the box." At the same time, we need to resolve unforeseen errors, difficulty, and other performance disturbances with task-specific measures ad hoc. Doing so requires domain-specific performance monitoring processes so as to improve upon and specialize task performance. Are there neural representations that support both domain-specific and domain-general performance monitoring, and if so, how are these requirements satisfied simultaneously?

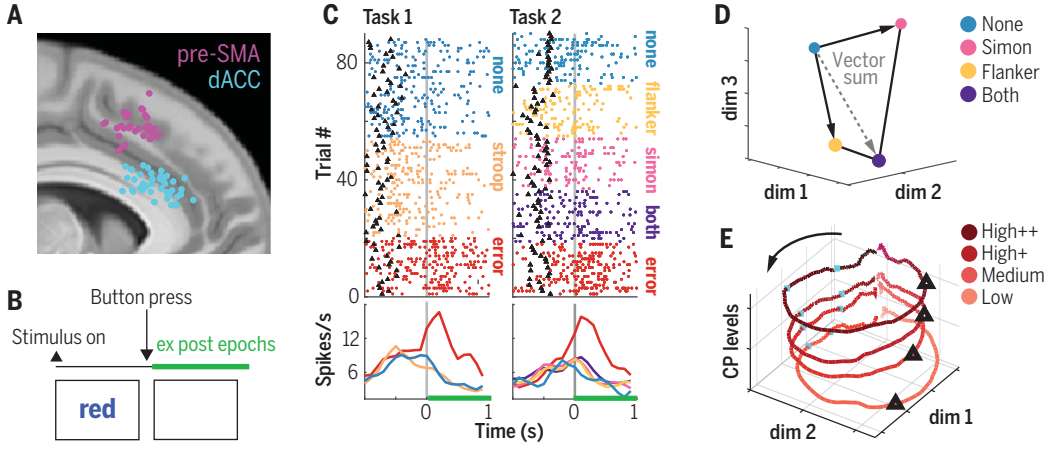
RATIONALE: In this work, we focused on the medial frontal cortex (MFC), which is known to serve a central role in performance monitoring. We recorded from single neurons in the MFC while neurosurgical patients performed two tasks in which errors were caused by different kinds of conflict. We identified neurons whose responses encoded performance monitoring variables in one or both tasks and characterized the resulting high-dimensional neural representations supported by neuronal populations. Although specialization and generalization place different constraints on the structure of neural representations, our work suggests that the geometry of population activity can be configured to accommodate both seemingly conflicting demands.

RESULTS: We recorded from 1431 single neurons in the MFC [dorsal anterior cingulate cortex (dACC) and pre–supplementary motor area (pre-SMA)] across 34 patients. Subjects were asked to press, as quickly as possible, one of three buttons to indicate the ink color of the word "red," "green," or "blue" (task 1, Stroop task) or to report the unique number among a string of numbers (task 2, multisource interference task). We modeled serial performance of both tasks using a hierarchical Bayesian framework. Our models assumed that subjects maintained an internal estimate of the probability of encountering a certain type of conflict (conflict probability, or CP) and iteratively updated this estimate using Bayes' law. Reaction times were generated by a drift-diffusion model, with CP biasing the drift rates. These behavioral analyses showed that the reaction times and likelihood of making an error varied systematically with trial history. At the single-neuron level, cells encoded different types of conflict, CP, and/or errors in one or both tasks. Individual neurons thus demonstrated heterogeneous task specificity when encoding these performance-monitoring variables, precluding a simplistic interpretation of domain generality at the single-neuron level. At the population level, however, these cells formed a high-dimensional representation with a geometry that allowed task-invariant decoding of all three variables on single trials, while at the same time allowing decoding of task-specific performance monitoring variables. Representations of conflict probability were consistent with that expected of the dynamics of a line attractor, with levels of CP stably maintained throughout the trial. The neural states indicating the presence of multiple types of conflict were equal to the vector sum of the states indicating each individual type of conflict, thereby revealing a compositional representation. Lastly, retrospective representations of conflict served to update internal estimates of conflict probability.

CONCLUSION: We leveraged the opportunity to record from the same populations of single neurons in the human MFC during two tasks to identify the structure of neural representations supporting performance monitoring. Population activity could be factorized into task identity and task-invariant dimensions that were orthogonal to each other. This geometry of the population activity could allow downstream brain regions to read out both domain-general and domain-specific signals from the same group of neurons and to initiate corresponding physiological and/or behavioral adaptations. These findings reveal how representations of evaluative signals can be both abstract and task-specific and suggest a neuronal mechanism for estimating control demand. ■

The list of author affiliations is available in the full article online.
*Corresponding author. Email: rutishauseru@csmc.edu (U.R.); zzbrooksfu@gmail.com (Z.F.)
Cite this article as Z. Fu *et al.*, *Science* **376**, eabm9922 (2022). DOI: 10.1126/science.abm9922

READ THE FULL ARTICLE AT
https://doi.org/10.1126/science.abm9922



Representations of evaluative signals in the human frontal cortex are both abstract and task-specific. (**A**) Recording locations. (**B**) Analysis epochs. (**C**) Response of example neuron in both tasks, demonstrating domain generality for errors (red) and no differentiation of different types of correct conflict trials (all other colors). (**D**) Compositionality of population-level conflict representations. (**E**) Conflict probability displaces dynamics in neural state space.

Downloaded from https://www.science.org at California Institute of Technology on May 05, 2022

RESEARCH ARTICLE

NEUROSCIENCE

The geometry of domain-general performance monitoring 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]*

Controlling behavior to flexibly achieve desired goals depends on the ability to monitor one's own performance. It is unknown how performance monitoring can be both flexible, to support different tasks, and specialized, to perform each task well. We recorded single neurons in the human medial frontal cortex while subjects performed two tasks that involve three types of cognitive conflict. Neurons encoding conflict probability, conflict, and error in one or both tasks were intermixed, forming a representational geometry that simultaneously allowed task specialization and generalization. Neurons encoding conflict retrospectively served to update internal estimates of conflict probability. Population representations of conflict were compositional. These findings reveal how representations of evaluative signals can be both abstract and task-specific and suggest a neuronal mechanism for estimating control demand.

Humans can rapidly learn to perform novel tasks given abstract rules, even if task requirements differ drastically. To achieve this, cognitive control must coordinate processes across a diverse array of perceptual, motor, and memory domains and at different levels of abstraction over sensorimotor representations (*1*–*3*). A key component of cognitive control is performance monitoring, which enables us to evaluate whether we have made an error, experienced conflict, and responded quickly or slowly (*4*, *5*). It provides task-specific information about which processes cause an error or a slow response so that they can be selectively guided (*6*–*16*). At the same time, performance monitoring needs to be flexible and domain-general to enable cognitive control for novel tasks (*17*), to inform abstract strategies [e.g., "win–stay, lose–switch," exploration versus exploitation (*18*, *19*)], and to initiate global adaptations (*20*–*28*). As an example, errors and conflicts can have different causes in different tasks (task-specific) but all signify failure or difficulty to fulfill an intended abstract goal (task-general); performance monitoring should satisfy both requirements. Thanks to its broad connectivity (*29*, *30*), the medial frontal cortex (MFC) serves a central role in evaluating one's own performance and decisions (*4*, *10*, *13*, *16*, *19*, *24*, *31*–*43*). However, little is known about how neural representation in the MFC can support both domain-specific and domain-general adaptations.

Specialization and generalization place different constraints on neural representations (*44*, *45*). Specialization demands separation of encoded task parameters, which can be fulfilled by increasing the dimensionality of neural representations (*46*, *47*). By contrast, generalization involves abstracting away details specific to performing a single task, which can be achieved by reducing the representational dimensionality (*3*, *46*). Theoretical work shows that the geometry of population activity can be configured to accommodate both of these seemingly conflicting demands (*44*), provided that the constituent single neurons multiplex task parameters nonlinearly (*48*, *49*). Although recent experimental work has shown that neuronal population activity is indeed organized this way in the frontal cortex and hippocampus in macaques (*44*) and humans (*50*) during decision-making tasks, it remains unknown whether this framework is applicable to the important topic of cognitive control.

A key aspect of behavioral control is learning about the identity and intensity of control needed to correctly perform a task and deploy control proactively on the basis of such estimates (*13*, *51*). This requires integrating internally generated performance outcomes over multiple trials. Blood oxygen level–dependent functional magnetic resonance imaging (BOLD-fMRI) studies localize signals related to control demand estimation to the insular-frontostriatal network (*52*). However, the neuronal mechanisms of how these estimates are updated trial by trial and whether the underlying substrate is domain-general or task-specific remain unknown.

Results

Task and behavior

Subjects (see table S1) performed the multi-source interference task (MSIT) and the color-word Stroop task (Fig. 1A and methods in the supplementary materials). Conflict and errors arose from different sources in these two tasks: competition between prepotency of reading and color naming in the Stroop task, and competition between the target response and either the spatial location of target ("Simon effect," denoted by "si") or flanking numbers ("flanker effect," denoted by "fl"), or both ("sf") in the MSIT. In the MSIT, we refer to trials with or without a Simon conflict as "Simon" and "non-Simon" trials, respectively (and similarly for flanker trials). Stimulus sequences were randomized, and each type of trial occurred with a fixed probability (see methods for details).

Subjects performed well (Stroop error rate: 6.3 ± 5%; MSIT error rate: 6.0 ± 5%). Reaction times (RTs) on correct trials were significantly prolonged in the presence of conflicts (Fig. 1B; fig. S1, D and E, shows raw RTs). Participants' successive performance (RT and accuracy) was modeled with a hierarchical Bayesian model (*52*–*54*). The model (Fig. 1C) assumes that participants iteratively updated internal estimates of how likely they were to encounter a certain type of conflict trial ("conflict probability") on the basis of their previous estimates and new evidence (experienced conflict and RT) on the current trial using Bayes' law. Given that trial sequences were randomized by the experimenter, subjects could not predict with certainty whether an upcoming trial would have conflict or not, but could instead estimate the probability of a conflict trial, which is fixed a priori but unbeknownst to the subject. The decision process was modeled as a drift-diffusion process (DDM), with drift rates being a function of conflict and subjects' estimated conflict probability ("CP coeff" in Fig. 1E). Model parameters (DDM parameters and trial-wise CP estimation) were inferred from subjects' behavioral data (RT and trial outcomes) and conflict sequences using data from all sessions and full Bayesian inference (see methods for details). Significance of model parameters was determined from the posterior distributions directly (see methods). The posterior predictive distributions captured the true RT distribution well (fig. S2, A and B). The Bayesian models were identifiable (fig. S2, C and D). Figure 1D shows RT and estimated conflict probability of an example MSIT session. Unless specified otherwise, we use the term "conflict probability" (or CP) to refer to

[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.
*Corresponding author. Email: rutishauseru@csmc.edu (U.R.); zzbrooksfu@gmail.com (Z.F.)
†Present address: Grossman School of Medicine, New York University, New York, NY, USA.

Downloaded from https://www.science.org at California Institute of Technology on May 05, 2022

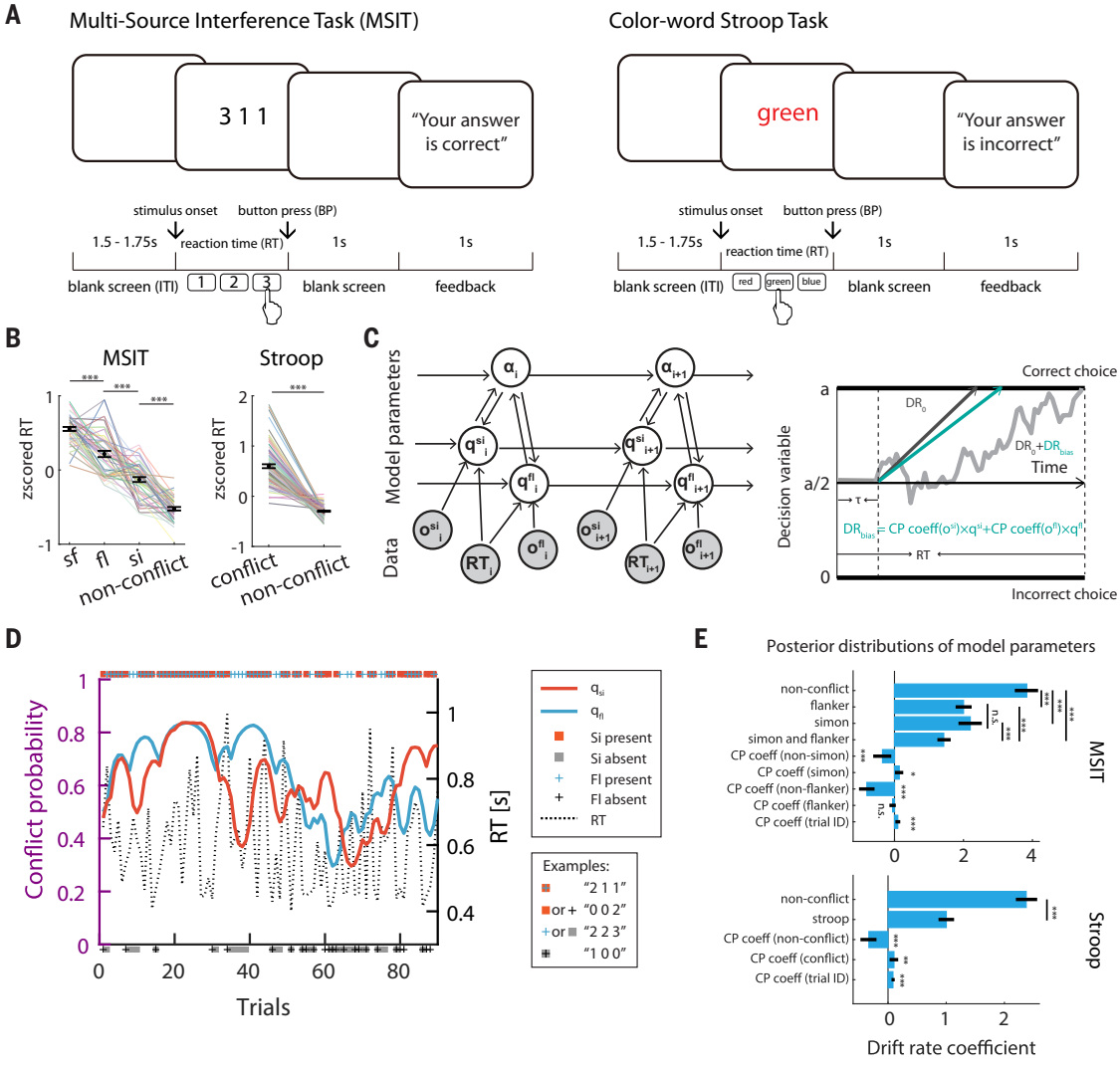


Fig. 1. Tasks, Bayesian conflict learning model, and reaction time analyses. (**A**) Task structure. (**B**) RTs were significantly prolonged by conflict in MSIT (left, n = 41 sessions) and Stroop task (right, n = 82 sessions). (**C**) The conflict probability estimation process (left) and the decision process modeled as a drift diffusion (right). Shown is the MSIT model, which has the six variables learning rate (α), Simon probability (q_{si}), flanker probability (q_{fl}), observed Simon conflict (o_{si}), observed flanker conflict (o_{fl}), and RT. Observables (trial congruency, RT, and outcome) are shown in gray; model parameters are in white. Arrows indicate information flow. (**D**) Estimated Simon probability (red) and flanker probability (blue) from an example MSIT session. Markers along the top indicate the type of conflict present. (**E**) Posterior distributions of model parameters after fitting to the behavior of all subjects. Black bars show high-density intervals. CP had a significant effect on RT. Vertical bars with asterisks show comparisons between posterior distribution, and asterisks without vertical bars mark comparisons with zero. *P < 0.05; **P < 0.01; ***P < 0.001; n.s., not significant (P > 0.05).

the posterior means of the estimated conflict probability on each trial.

Conflicts reduced drift rates and thus prolonged RT [$P(dr_{\text{Stroop nonconflict}} < dr_{\text{Stroop conflict}}) < 0.001$, $P(dr_{\text{MSIT nonconflict}} < dr_{\text{Simon}}) < 0.001$, $P(dr_{\text{MSIT nonconflict}} < dr_{\text{flanker}}) < 0.001$, $P(dr_{\text{MSIT nonconflict}} < dr_{\text{Simon and flanker}}) < 0.001$]. High estimated CP prolonged RT on non-conflict trials [negative CP coefficient decreases drift rate: $P(CP\ coeff_{\text{Stroop nonconflict}} > 0) < 0.001$, $P(CP\ coeff_{\text{non-Simon}} > 0) < 0.001$, $P(CP\ coeff_{\text{nonflanker}} > 0) < 0.001$] but hastened RT on conflict trials [positive CP coefficient increases drift rate: $P(CP\ coeff_{\text{Stroop}} < 0) < 0.01$, $P(CP\ coeff_{\text{Simon}} < 0) < 0.05$] (Fig. 1E). Our model yielded qualitatively similar results when applied to data from a separate group of healthy control subjects (fig. S1C). In a separate control model where only trial outcomes and conflict sequences were used (but not RT), high CP correlated with reduced likelihood of making an error on conflict trials [Bayesian logistic regression; negative CP coefficient reduces error likelihood: $P(CP\ coeff_{\text{Stroop}} > 0) < 0.01$, $P(CP\ coeff_{\text{Simon}} > 0) < 0.05$] (fig. S1B), suggesting that higher CP predicted higher levels of control. See (*55*) for more about model comparisons.

Neuronal correlates of performance-monitoring signals

We collected single-neuron recordings from two regions within the MFC (Fig. 2A): the dorsal anterior cingulate cortex (dACC) and the pre–supplementary motor area (pre-SMA). We isolated 1431 putative single neurons [Stroop: 584 in the dACC and 607 in the pre-SMA across 34 participants (10 females); MSIT: 326 in the dACC and 412 in the pre-SMA in 12 participants (6 females)]. Neurons in the dACC and the pre-SMA were pooled, unless specified otherwise, because neurons responded similarly (but see the section "Comparison between the dACC and the pre-SMA" below for notable differences).

We identified neurons selective for the mean and variance of the posterior distribution of CP, error, and conflict surprise (unsigned prediction error of CP: 1 – |*conflict* – *CP*|, where *conflict* is an indicator variable) in the ex post epoch, and conflict in the ex ante and ex post epochs (see example neurons in Fig. 2, schematic of analysis epochs in Fig. 3A, and a summary of overall cell counts in Fig. 3B).

A substantial proportion of neurons encoded the mean and variance of CP posterior distribution in the ex post epoch (Fig. 3B; 25 and 17% of neurons in MSIT, 21 and 12% in Stroop). CP is maintained up to the baseline of the next trial [Fig. 3B, "conflict prob. (baseline)", blue; 21% in MSIT, 17% in Stroop]. Neurons encoded conflict in the ex ante epoch (17 and 14% of neurons in MSIT and Stroop, respectively;

Downloaded from https://www.science.org at California Institute of Technology on May 05, 2022

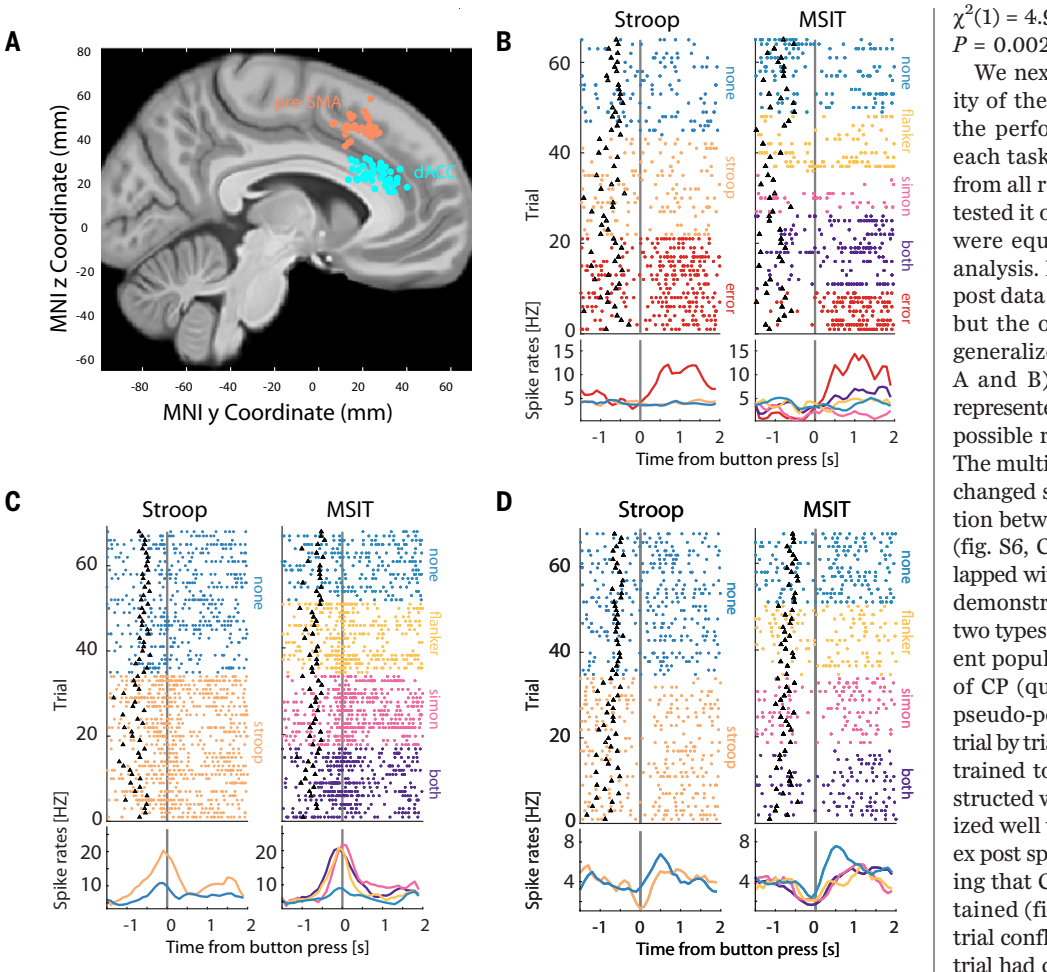


Fig. 2. Recording locations and example neurons. (**A**) Recording locations shown on the CIT168 Atlas brain. Each dot indicates the location of a microwire bundle. MNI, Montreal Neurological Institute. (**B** to **D**) Activity of three example neurons that show similar response dynamics in both tasks: neuron signaling action error (B), conflict by firing rate increase (C), and conflict by firing rate decrease (D). The black triangles mark stimulus onset. Trials are resorted by type and subsampled to equalize trial numbers for visualization only.

Fig. 3B, cyan), consistent with previous reports (*33*, *38*), but, intriguingly, also in the ex post epoch (21% in MSIT, 20% in Stroop). Neurons encoded conflict surprise (16% in MSIT, 12% in Stroop) and errors (20% in MSIT, 30% in Stroop; see Fig. 2B for an example) in the ex post epoch. The distributions of selective units were similar between the two tasks (Fig. 3B).

Approximately 30% of conflict neurons were selective exclusively in either the ex ante, early (0 to 1 s after button presses), or late (1 to 2 s after button presses) ex post epochs (Fig. 3C), with some (18%) selective in the ex ante, early, and late ex post epochs ("extended"). Distributions of conflict signals across time were similar between the MSIT and Stroop tasks (Fig. 3C). We were intrigued by neurons signaling conflict ex post (20% of neurons in both tasks; Fig. 2, C and D, shows examples), which has not been reported before. This conflict signal, which arises too late to be useful for within-trial cognitive control, was more prominent than the one found in the ex ante epoch in both tasks [15% versus 20%; $\chi^2(1) = 18.34$, $P < 0.001$, chi-squared test].

Many conflict neurons also signaled errors, surprise, CP, or a conjunction of these variables (Fig. 3D). This multiplexing depended on the timing of conflict signals. The proportion of conflict neurons that multiplexed CP (light green bars) increased significantly toward the end of the trial, when updating would be mostly complete [proportion in the late ex post epoch versus that in all other epochs; $\chi^2(1) = 6.86$, $P = 0.008$ for MSIT; $\chi^2(1) = 3.6$, $P = 0.04$ for Stroop, chi-squared test]. Consistent with this idea, the group of neurons signaling conflict exclusively in the ex ante epoch showed the least multiplexing, consistent with a primary role in monitoring conflict during action production [proportion of "pure" conflict neurons active only during the ex ante epoch versus those that are active in other epochs; $\chi^2(1) = 4.93$, $P = 0.03$ for MSIT; $\chi^2(1) = 9.36$, $P = 0.002$ for Stroop, chi-squared test].

We next investigated the temporal stability of the multivariate coding patterns for the performance monitoring variables in each task. We trained a decoder with data from all recorded neurons in one epoch and tested it on data from a different epoch. RTs were equalized across conditions for this analysis. Error decoders trained on early ex post data generalized poorly to later epochs, but the ones trained on late ex post data generalized well into earlier epochs (fig. S6, A and B), suggesting that some neurons represented errors only late in the trial, with a possible role in post-error adjustments (*33*). The multivariate coding patterns for conflict changed substantially to prevent generalization between the ex post and ex ante epochs (fig. S6, C to E; green horizontal lines overlapped with orange and blue only minimally), demonstrating that the MFC encoded these two types of conflict information with different populations of neurons. The magnitude of CP (quantized to allow construction of pseudo-populations) could also be decoded trial by trial from population activity. Decoders trained to decode CP before updating (constructed with baseline spike counts) generalized well to decode CP after updating (using ex post spike counts), and vice versa, suggesting that CP representation was stably maintained (fig. S7). Representation of previous trial conflict (indicator; "1" denotes previous trial had conflict) was relatively weak (fig. S8, A to C; <60% decoding accuracy), consistent with the model comparison results where previous conflict alone was a poor predictor of RT compared with estimated CP (tables S2 and S3). CP neurons demonstrated the greatest degree of multiplexing (Fig. 3F). Only 18% of these neurons signaled CP exclusively, with the remainder multiplexing information about estimated CP on the last trial, conflict, conflict surprise, or a conjunction of these variables. CP neurons changed their firing rates ex post by an amount commensurate with the numerical change of conflict probability between contiguous trials (Fig. 3G, $P < 0.001$, t test against zero; mean correlations in Simon, flanker, and Stroop are 0.036, 0.036, and 0.045, respectively), reflecting a neural "updating" process. Although for some neurons the updating process started in the ex ante epoch, for the majority it started in the ex post epoch and continued well into the ex post period (fig. S3, C to E, shows time courses).

CP neurons differed qualitatively from those of other recorded neurons. First, their trial-by-trial baseline spike rates (treated as a time series) had higher self-similarity than all other types of neurons studied [fig. S4, A and B; self-similarity quantified by α values from the detrended fluctuation analysis (DFA); fig. S4, C to F,

Downloaded from https://www.science.org at California Institute of Technology on May 05, 2022

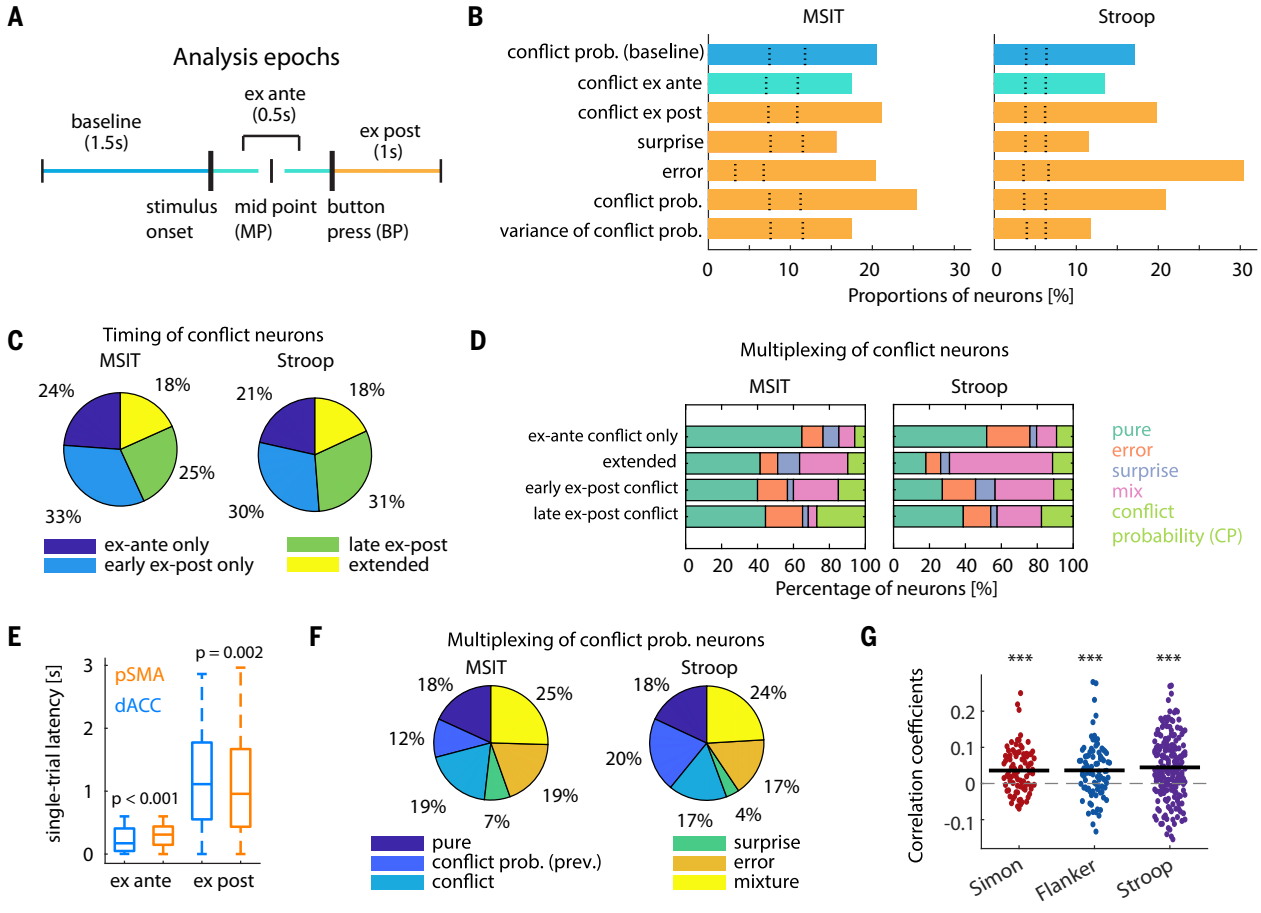


Fig. 3. Single-neuron tuning properties. (**A**) Illustration of epochs used for analysis. Thick vertical bars represent physical events, the slim vertical bars demarcate epochs. (**B**) Percentage of neurons encoding the variable indicated in the two tasks. The color code is as indicated in (A). Dotted lines represent 2.5th and 97.5th percentiles of the null distribution obtained from permutation. For all groups shown, $P < 0.001$. (**C**) Percentage of conflict neurons that are selective in each time period. Early and late ex post epochs denote 0 to 1 s and 1 to 2 s after button press, respectively. (**D**) Percentage of conflict neurons that were also selective for error, surprise, CP, or any combination of these factors ("mix"). (**E**) Comparison of single-trial neuronal response latency of conflict neurons in dACC and pre-SMA (correct trials only, $t = 0$ is stimulus onset for ex ante and button press for ex post conflict neurons). (**F**) Percentage of conflict probability neurons that were also selective for CP on the last trial, conflict, surprise, error, or all combinations of these variables ("mix"). (**G**) Neuronal signature of updating conflict probability estimation. Correlation is computed between the difference between current estimation and conflict probability on the last trial (behavioral update) and the difference between demeaned $FR_{ex\ post}$ and $FR_{baseline}$ (neural update) for all conflict probability neurons. ***$P < 0.001$.

shows DFA analysis of two example neurons]. Second, the extracellular spikes generated by these neurons had significantly narrower widths compared with other neurons (fig. S4, G and H, right panels; more broadly, DFA α is negatively correlated with spike width, shown in fig. S4, G and H, left panels).

Representational geometry of conflict within a single task

We next investigated the representational geometry of conflict in the high-dimensional neural state space formed by all recorded neurons. This is possible because different types of cognitive conflict coexist within MSIT. We first tested whether the four MSIT conflict conditions (si, fl, sf, and nonconflict) are separable in the neural state space. A "coding dimension" was defined as the high-dimensional vector flanked by the means of the sf and nonconflict ("none") trials (Fig. 4A, broken lines; see methods). Held-out single-trial data projected onto this coding dimension allowed a decoder to differentiate with high accuracy between all pairs of conflict conditions in the ex ante epoch (Fig. 4B, left), and all but one pair (si versus sf, $P = 0.16$) in the ex post epoch (Fig. 4B, right). Notably, si and fl conflict could be differentiated above chance by this coding dimension. Conflict conditions were still separable after equalizing for RT across conditions (fig. S9C; see methods for RT equalization), suggesting that the separation was independent of trial difficulty for which RT is a proxy (*56*). We next investigated whether representations generalized between Simon and flanker conflict. For each time bin, a coding dimension was constructed by connecting, in the neural state space, the mean of Simon (si + sf) trials with the mean of nonSimon (fl + none) trials, and a separate one by connecting the flanker (fl + sf) versus nonflanker (si + none) means. Held-out single-trial data from either conflict type projected onto the coding dimension constructed using data from another conflict type allowed decoding with high accuracy (Fig. 4C; gray trace, flanker coding dimension decoding Simon versus non-Simon; black trace, Simon coding dimension decoding flanker versus nonflanker).

We next tested whether the representation of conflict was compositional. If true, relative to the mean of "none" trials, the sf representation should be located where the linear vector sum of Simon (si) and flanker (fl) trials lands, forming a parallelogram with the none-sf axis at its diagonal (Fig. 4A). To test this with a decoding approach, a decoder trained to differentiate between the two classes connected by one edge of this parallelogram should be

Downloaded from https://www.science.org at California Institute of Technology on May 05, 2022

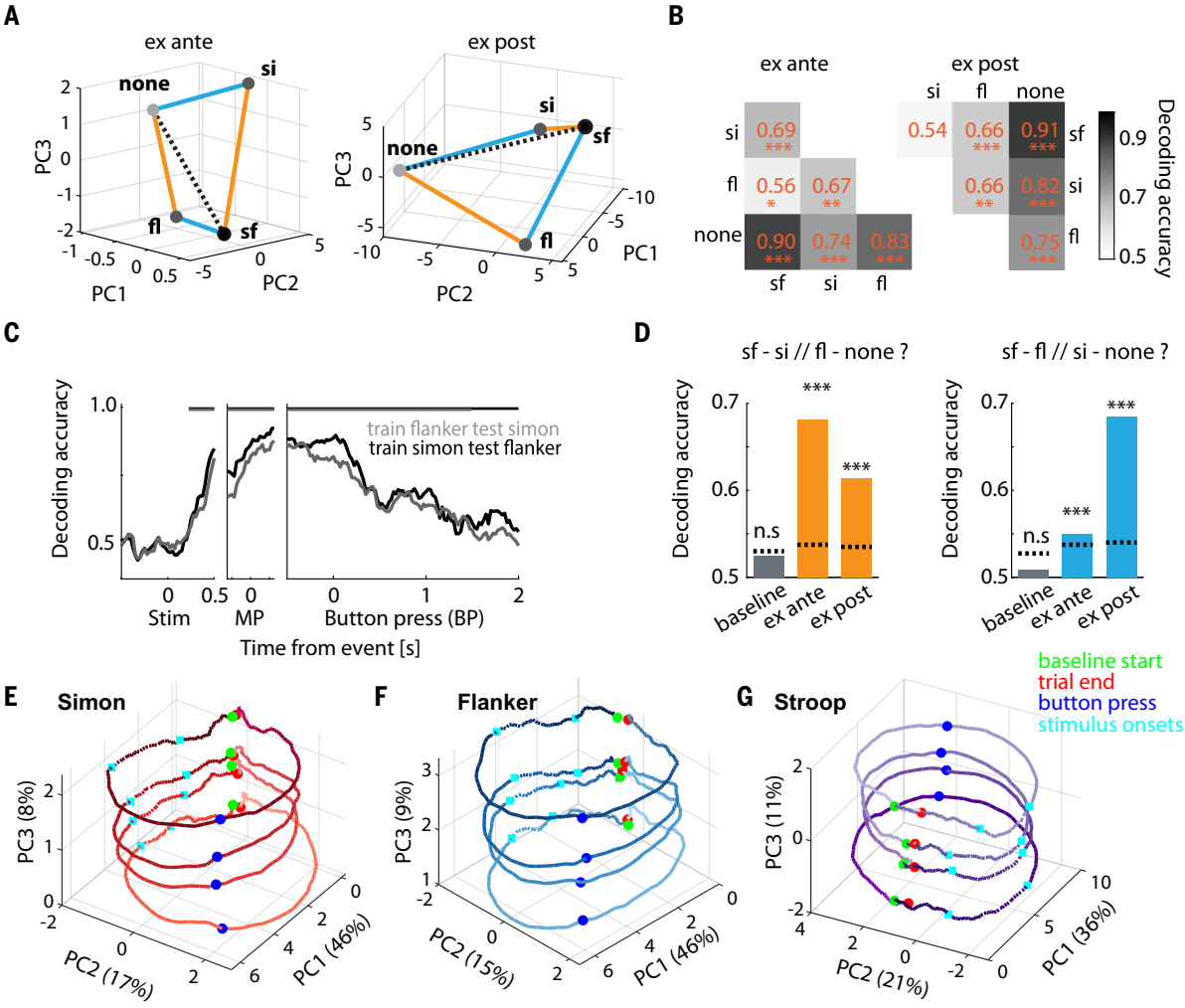


Fig. 4. State-space representation of conflict and conflict probability. (**A**) State-space representation of conflict (left, ex ante; right, ex post) in MSIT task visualized in PCA space. Dotted line is the vector used to classify pairs of conflict conditions in (B). (**B**) Decoding accuracy from classification of pairs of conflict conditions in MSIT. (**C**) Coding dimensions invariant between Simon and flanker conflict. At each time point, a decoder is trained on Simon versus non-Simon and tested on held-out flanker versus nonflanker trials (black) and vice versa (gray). (**D**) Conflict representations are compositional. Decoders trained on one edge of the parallelogram were able to differentiate between conditions along the opposite parallel edge [orange and blue edges shown in (A), respectively, shown on left and right]. Dotted lines show 97.5th percentile of the null distribution. (**E** to **G**) State-space representation of CP in MSIT and Stroop, visualized in PCA space. Green dots mark baseline start, the two cyan squares delineate the range of stimulus onsets, blue dots mark button press, and red dots mark end of trial. Trials are aligned to button press. Color fades as the trial progresses. Numbers signify percentage of variance explained by each principal component (PC). *P < 0.05; **P < 0.01; ***P < 0.001; n.s., not significant (P > 0.05). MP, midpoint; BP, button press; Stim, stimulus onset.

able to decode the two classes connected by the opposite edge (and vice versa). Indeed, a decoder trained to differentiate sf from fl trials, which is simply the vector connecting sf and fl (blue edge in Fig. 4A), was able to decode si from nonconflict trials projected to this axis with above-chance performance and vice versa (Fig. 4D, P < 0.001 for both the ex ante and ex post data, permutation test). The same was true for the other pair of edges (Fig. 4D, testing blue edges in Fig. 4A; P < 0.001 for both the ex ante and ex post data, permutation test). The parallelism was imperfect because the decoding accuracy, while well above chance, was relatively low (<70%) compared with within-condition decoding performance (Fig. 4C). Neurons that demonstrated nonlinear mixed selectivity for Simon and flanker conflicts (as measured by the F statistic of the interaction term between Simon and flanker derived from an analysis of variance model) contributed the most to this deviation from parallelism at the population level [fig. S9F, correlation coefficient (r) = 0.74, P < 0.001, for ex post data; fig. S9E, r = 0.75, P < 0.001, for ex ante data; Spearman's rank correlation). This representation structure was disrupted on error trials: generalization performance dropped in the ex ante (fig. S9D; for both edges, 68 and 61% on correct trials versus 56 and 47% on error trials) as well as the ex post epoch (fig. S9D; for both edges, 55 and 68% on correct trials versus 51 and 59% on error trials) on error trials.

Representational geometry of conflict probability

Conflict probability representation can be viewed as a state (an initial condition) that is present before stimulus onset and to which the population returns after completing a trial. To test this idea, we binned trials by CP quartiles (four "levels" were generated as trial labels) for each session and aggregated neuronal data across different sessions to generate pseudopopulations. Principal components analysis (PCA), performed separately for each type of conflict, revealed that the variability across

Downloaded from https://www.science.org at California Institute of Technology on May 05, 2022

different CP levels (8 to 11% of explained variance) was captured mostly by a single axis (PC3s in Fig. 4, E to G; green dots mark trial start, red dots mark trial end), which was orthogonal to time-dependent but condition-independent firing rate changes (captured by PC1s and PC2s, 57 to 63% of explained variance). We quantified this impression in the full neural state space (not in PCA space). Before stimulus onset (demarcated by green and cyan filled circles in Fig. 4, E to G), the neural state evolved with low speed (blue points in fig. S9, G to I). The neural state then changed more quickly after stimulus onset (orange points in fig. S9, G to I; $P < 0.001$, paired t test), before eventually returning to a low-speed state that hovered near the starting position (red filled circles in Fig. 4, E to G; yellow points in fig. S9, G to I). The distance between the four trajectories was also kept stable across time. A coding dimension in the full neural space (which corresponds to PC3 in Fig. 4, E and F) was constructed by connecting the means of neural representations of the first and fourth CP levels. The held-out single-trial data, when projected onto this coding dimension, separated into four levels with their ordering (in projection values) being strongly correlated with the ordering of CP levels (fig. S9, J to L). This ordering is maintained into the baseline period of the next trial (fig. S9, J to L). For single sessions, CP predicted using spike counts from all MFC neurons in the session as regressors were correlated with the true values of CP trial by trial (fig. S15, M and N). These analyses in the full neural space confirmed the insights gained from PCA visualization in Fig. 4, E to G.

Domain-general performance-monitoring signals at the population level

We next investigated whether the geometry of performance-monitoring representations supported readouts that are invariant across MSIT and Stroop, while simultaneously allowing robust separation of conditions specific to MSIT. Within MSIT, a geometry can be extracted that supports invariance across types of conflicts while keeping the four conflict conditions separate (Fig. 4, A to C). However, it is unclear whether such representation is specific to a single task and requires construction of a task set. We thus studied the activity of the same neurons in two behavioral tasks performed separately (table S1 shows session information). We used demixed PCA (dPCA) (*57*) to factorize population activity into coding dimensions for performance monitoring variables (error, conflict, and conflict probability) and for the task identity. The dPCA provided a principled way to optimize for cross-task decoding and could serve as an existence proof for domain generality. The statistical significance of the dPCA coding dimensions was similarly assessed by out-of-sample decoding. Successful "demixing" should correspond to (near) orthogonality between the performance monitoring and task identity coding dimension.

The dPCA coding dimensions extracted using error and conflict contrasts (Stroop-Simon and Stroop-flanker separately) each explained 10 to 21% of variance, supported task-invariant decoding across time, and were orthogonal to the task identity dimension (see Fig. 5A for error, Fig. 5B for Stroop-Simon, fig. S10A for Stroop-flanker conflict decoding over time, and Fig. 5C for conflict decoding restricted in the ex ante and ex post epochs separately; for all clusters, $P < 0.01$, cluster-based permutation tests; clusters with significant decoding performance demarcated by horizontal bars; for error, angle = 92.1°, $P = 0.92$, $\tau = 0.005$; for Stroop-Simon conflict, angle = 81.9°, $P = 0.31$, $\tau = 0.04$; for Stroop-flanker conflict, angle = 80.8°, $P = 0.09$, $\tau = 0.06$; Kendall's rank correlation).

The task-invariant conflict coding dimension (extracted by removing the task difference between MSIT sf and Stroop conflict trials and between nonconflict trials in both tasks) was also able to differentiate five out of six pairs of conflict conditions with high accuracy (60 to 90%) within MSIT in both the ex ante and ex post epochs (Fig. 5C; for statistics, see figure, permutation tests). Similar task-invariant decoding performance was obtained when restricting to only trials with similar RTs across conditions, suggesting again that task-invariant representations of error and conflict were not due to subjective difficulty (fig. S10, B to E), for which RT is a proxy (*56*).

Similarly, the coding dimensions for Stroop-Simon conflict probability (Fig. 5D and fig. S10G) and Stroop-flanker conflict probability (fig. S10, F and H) each explained 27 to 35% of variance, supported task-invariant decoding, and were orthogonal to the task identity dimension ($P > 0.05$, Kendall's rank correlation). Notably, this task generalizability did not compromise the capacity of this coding dimension to separate different levels of CP within Stroop or MSIT (Fig. 5D and fig. S10, F to H).

As a control analysis, we also examined whether a decoder trained on a single task would generalize to the other task. These "within-task" coding dimensions supported decoding of data from the other task with high accuracy, albeit with slightly lower performance than the dPCA coding dimensions that were optimized for cross-task decoding (fig. S12, A to F, for decoding performance, weight correlation between tasks). This is expected given the robust correlation between the weights of these within-task coding dimensions (scatters in fig. S12, G to M). This correlation was confirmed by the fact that the angles between the two within-task coding dimensions were significantly smaller than the null distribution (which approximately equaled 90°; inset histograms in fig. S12, G to M). The angle between the task-invariant coding dimensions optimized by dPCA and the within-task dimensions trained with either Stroop or MSIT data alone was similar (angularly "equidistant"; angles reported in insets, fig. S12, G to M). The optimization performed by dPCA can thus be understood geometrically: It sought a coding dimension that is angularly "centered" between each of the within-task dimensions trained using data from one task only. Task-invariant representations for all performance monitoring variables can be found using simultaneously recorded data at the single-session level in subjects with enough neurons (fig. S15), highlighting that it is the same population of neurons that subserves such a geometry.

Subjects responded more slowly on the trials that followed error trials in both tasks, demonstrating robust post-error slowing (PES: 40 ± 1 ms and 44 ± 1 ms in Stroop and MSIT, respectively; fig. S8D). However, this trial-by-trial measure of PES was not significantly correlated with spike rates (on the error trials) of error neurons in either task individually, or for error neurons that signaled errors in both tasks (see methods for definition of "task-invariant"; fig. S8E).

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

We sought to understand how tuning profiles of single neurons contributed to the population geometry that simultaneously allowed task-invariant and task-specific readouts. While some neurons encoded error, conflict, and CP in a task-invariant manner (Figs. 2 and 3), others encoded these variables in only one task. The extent to which single-neuron tuning depended on the task was assessed using partial correlations ("task-invariant turning strength"; higher correlation with variable of interest after controlling for task variable means higher task invariance). According to this test, neurons that had significant correlation were labeled "task-invariant" (65 to 83%), and those with insignificant cross-task correlation but significant correlation within either task were labeled "task-dependent" (17 to 35%; Fig. 5, E and F, and fig. S13, red and cyan slices, respectively). Across all neurons, the dPCA weight assigned to each neuron was strongly correlated with the neuron's task-invariant tuning strength (scatter plots in Fig. 5, E and F, and fig. S13; $P < 0.001$ for all panels, Spearman rank correlation). Both the task-invariant and the task-dependent neurons were assigned significant larger weights compared with nonselective ("other") neurons (Fig. 5, E and F, and fig. S13, dot density plots). Similar conclusions hold when using a

Downloaded from https://www.science.org at California Institute of Technology on May 05, 2022

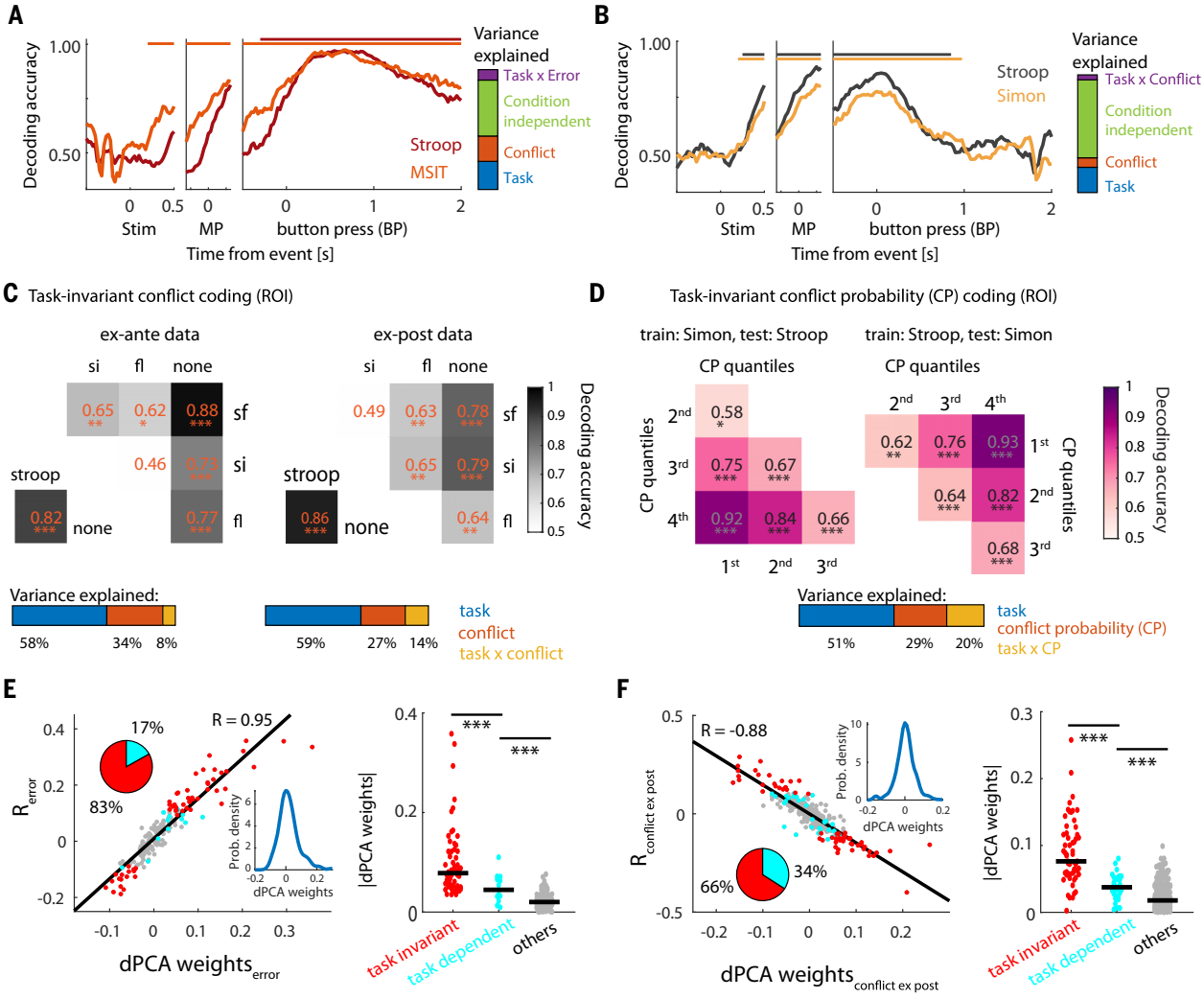


Fig. 5. Domain-general representation of performance-monitoring signals. (**A**) Task-invariant decoding of errors. Bar on the right shows the variance explained by the different dPCA components. Data from the whole trial were used. (**B**) Task-invariant decoding of conflict. The bar on the right represents variance explained by the different dPCA components (see figure legend for color code). Data from the whole trial were used. (**C**) Separability of conflict conditions along the domain-general conflict axis in both the ex ante (left) and ex post epochs (right). The dPCA coding dimension was constructed (separately for each epoch) using Stroop, sf conflict, and nonconflict trials and supported decoding of Stroop, Simon, and flanker conflicts ("task-invariant") as well as separation of Simon and flanker conditions. (**D**) Task-invariant decoding of CP. The dPCA coding axis was constructed using Stroop and Simon CP (binned by quartiles into four levels) and supported pairwise decoding of CP levels in both tasks. For (C) and (D), horizontal colored bars at the bottom show variance explained of dPCA coding dimensions. Data from single regions of interest (ROIs) were used. (**E** and **F**) Relationship between task-invariant single-neuron tuning strength of error (E), conflict ex post (F), and the corresponding dPCA weights. Pie charts show the percentages of task-invariant neurons (red slice) that had a significant main effect for the performance monitoring variable and those that had significant effect only in either MSIT or Stroop ("task-dependent", blue slice). Scatter plots (left) show significant correlation between task-invariant coding strength and the corresponding dPCA weights. The *y* axis shows correlation of firing rate of a neuron with the given variable, after removing task information by partial correlation (see methods). *P < 0.05; **P < 0.01; ***P ≤ 0.001; n.s., not significant (P > 0.05).

permutation-based feature-importance measure instead of the dPCA weights (fig. S14; see methods for details). Notably, the distributions of dPCA weights or feature importance (probability density shown in insets in Fig. 5, E and F, and figs. S13 and S14) across neurons, with many showing nonzero weights, suggested that the task-invariant coding was not driven by few neurons but instead reflected population activity.

Neurons with higher self-similarity for baseline spike counts contributed more to the task-invariant coding dimension of conflict probability after controlling for single-neuron coding strength (partial correlation between absolute values of dPCA weights and DFA α values; r = 0.16, P = 0.003 for Stroop-Simon conflict probability; r = 0.20, P < 0.001 for Stroop-flanker conflict probability).

Comparison between the dACC and the pre-SMA

The distributions of neuronal selectivity were similar in dACC and pre-SMA (fig. S3, A and B). Both areas contained similar proportions of task-invariant and task-dependent neurons. Both areas independently supported compositional conflict coding and domain-general readouts for error (fig. S11, A and B), conflict (fig. S11, C to F), and conflict probability (fig. S10, G to J). However, there were three notable differences between the two areas. First, the temporal profiles of decoding performance for error, conflict, and conflict probability were similar between the areas, but decoding accuracy in pre-SMA was consistently higher (fig. S5, A to E). Second, the task-invariant error

Downloaded from https://www.science.org at California Institute of Technology on May 05, 2022

response appeared earlier in pre-SMA than in dACC ($\Delta_{latency}$ = 0.55 s for Stroop and 0.5 s for MSIT), consistent with our previous report in the Stroop task (*33*). Third, ex ante conflict information was first available in dACC, followed by pre-SMA (Fig. 3E; median difference = 138 ms; $P < 0.001$, Wilcoxon rank sum test; single-trial spike train latency). Third, by contrast, ex post conflict information was available first in pre-SMA, followed by dACC (Fig. 3E; median difference = 161 ms; $P = 0.002$, Wilcoxon rank sum test).

Discussion

Within the MFC, some neurons encoded error, conflict, and conflict probability in a task-invariant way; some encoded these variables exclusively in one task; and others multiplexed domain and task information to varying degrees. They were intermixed at similar anatomical locations and demonstrated nonlinear "mixed selectivity" (*47*, *48*) for task identity and performance monitoring variables. Such complexity at the single-neuron level precludes a simplistic interpretation of domain generality. By contrast, population activity can be factorized into a task dimension and a task-invariant dimension that were, in most cases, orthogonal to each other. The geometry of MFC population representation allows simple linear decoders to read out performance monitoring variables with high accuracy (>80%) in both tasks, and simultaneously to differentiate different types of conflict or different levels of estimated conflict probability within a task. Notably, it is the same group of neurons that gave rise to this geometry. This finding contrasts with neuroimaging studies that either report domain-specific and domain-general conflict signals encoded by distinct groups of voxels (*58*, *59*) or an absence of domain-general conflict signals (*60*) within the MFC.

Several studies have proposed that the lateral prefrontal cortex (LPFC) is topographically organized to subserve cognitive control, with more abstract processing engaging the anterior regions (*1*–*3*, *61*, *62*). Unlike in the LPFC, domain-general and domain-specific neurons were intermixed within the MFC. The representational geometry we report here is well suited to provide performance-monitoring signals to these subregions: Downstream neurons within these LPFC regions can select performance-monitoring information at different degrees of abstraction by adjusting connection weights, similar to input selection mechanisms described in the PFC (*63*). Curiously, we did not observe a prominent difference between the pre-SMA and the dACC in the degree of domain generality of performance monitoring, which may be related to the fact that these two areas are highly interconnected.

Domain-general error signals

A key component of performance monitoring and metacognitive judgment is the ability to self-monitor errors without relying on external feedback (*4*, *37*, *40*, *41*, *64*, *65*). A subset of neurons signaled error not only in the Stroop task [as previously reported in (*33*)] but also in MSIT. The error signal is thus domain-general, that is, abstracted away from the sensory and motor details as well as the types of response conflicts encountered across the two tasks (Fig. 2B). At the population level, these domain-general error neurons enabled trial-by-trial readout of self-monitored errors with >90% accuracy equal across tasks (Fig. 5, A and E). Compatible with earlier results (*33*), the activity of neither task-specific nor task-general error neurons directly predicted the extent of PES. This result is consistent with the evaluative roles of MFC error signals and suggests that the control process may lie outside of the MFC. Given the fMRI-BOLD finding that the MFC is a domain-general substrate for metacognition (*66*), an interesting open question is whether the same neural mechanisms we describe here support metacognitive judgment across different domains, such as perceptual or memory confidence.

Domain-specific performance monitoring

The causes of conflict and errors in the two tasks differ: distraction by the prepotent tendency to read in the Stroop task, and distraction by location of target number (Simon) or by numbers flanking the target (flanker) in the MSIT. Such performance perturbations call for specific compensatory mechanisms, such as suppressing attention to task-irrelevant stimulus dimensions. Consistent with this requirement, a subset of neurons signaled errors, conflict, and conflict probability exclusively in one task, giving rise to a task identity dimension that supported robust decoding of which task the performance disturbance occurred in (>90% accuracy). This provides the domain-specific information about the sources of performance disturbances for cognitive control, consistent with the reported role of MFC neurons in credit assignment (*16*). The existence of task-specific neurons also suggests that the performance monitoring circuitry can be rapidly and flexibly reconfigured in different tasks to subserve different task sets (*67*, *68*), consistent with the rapid reconfiguration of functional connectivity among cognitive control networks to enable novel task performance (*69*).

Compositionality of conflict representation

Compositionality of conflict representation can be formulated as a problem of generalization: If Simon and flanker conflict are linearly additive, decoders trained to identify the presence of only Simon or flanker conflict should continue to do so when both types of conflict are present. We found that this was the case, with the neural state approximately equal to the vector sum of the two neural states when the two types of conflict are present individually. 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

Our model for conflict probability estimation predicts that conflict should be signaled twice: once during response competition (ex ante) and again after the action has been committed (ex post). Whereas the former, predicted by conflict monitoring theory (*13*, *70*), provides transitory metacognitive evidence of conflict and is important for recruiting within-trial cognitive control (*70*), the latter provides a stable "indicator" of whether conflict occurs or not. Both signals arise independently of external feedback, thus qualifying as correlates of metacognitive self-monitoring (*64*, *65*). One interpretation for the ex post signals (conflict or error), as proposed by connectionist models of conflict monitoring, is that they reflect conflict between the committed response and continuing stimulus processing (*70*, *71*), which should also activate the ex ante conflict neurons. However, our data demonstrated that these two types of conflict signals were encoded by separate neurons and their multivariate coding patterns differed substantially enough to prevent generalization (fig. S6, C to E).

There is significant overlap between error neurons and ex post conflict neurons. Common coding dimensions that simultaneously decode both error and conflict do exist, although the decoding accuracy for conflict is significantly lower than for error (fig. S9, A and B). In macaque SEF, ex post activation is found after noncanceled and successfully canceled saccades (*41*). These ex post evaluative signals may reflect a common process that compares a corollary discharge signal (conveying the actual state of action selection and cognitive control) with cognitive control state predicted by "forward models" (*72*–*74*) and may underlie sense of agency (*75*). Future work is needed to test this hypothesis.

The neurons that reported conflict probability changed their firing rates trial by trial, and this "updating" in firing rates occurred primarily upon commitment of an action.

Downloaded from https://www.science.org at California Institute of Technology on May 05, 2022

The distinct functional properties of conflict probability neurons, that is, significantly narrower extracellular spike waveforms and higher self-similarity of baseline spike counts (fig. S4, C to H), suggest that they may be interneurons with strong recurrent connectivity, consistent with a prior report where dACC neurons that encode past outcomes have narrower extracellular waveforms (*76*). At the population level, these single-neuron properties contribute to the formation of task-invariant line attractor dynamics. Computational modeling demonstrates the importance of inhibitory interneurons in maintaining information in working memory, which occurs on the scale of several seconds (*77*). Similar circuit-level mechanisms could provide a basis for retaining the history of performance monitoring and reward, which is on the scale of several minutes. These conflict probability neurons might thus provide a neural substrate for proactive control and learning the value of control, both of which require stable maintenance of learned information (*13*, *51*). The trade-off between flexible updating and stable maintenance of performance monitoring information remains an open question.

Materials and methods summary

Detailed materials and methods with references can be found in the supplementary materials. Briefly, the subjects were 34 patients who were undergoing intracranial monitoring of epileptic seizures using hybrid depth electrodes with embedded microwires. Spikes were detected and sorted using a template-matching algorithm, and only well-isolated single neurons were analyzed. Electrodes were localized using postoperative imaging, and all included subjects had well-isolated neuronal activity on at least one dACC or pre-SMA electrode.

Subjects performed speeded version of the Stroop and MSIT tasks. For the Stroop task, subjects were instructed to name the color in which the word shown on the screen (red, green, or blue) was printed (red, green, or blue). For the MSIT task, subjects were shown three numbers (0, 1, 2, and/or 3), out of which two were the same and the third was different (target), and were instructed to report the target number identity. After each response, the stimulus screen was immediately blanked (1 s), followed by a feedback screen ("correct," "incorrect," or "too slow"; 1 s). Trial sequences were randomized.

We used a hierarchical Bayesian model, which iteratively updates an internal estimate of conflict probability trial by trial. Parameters were the learning rate α and conflict probability (q_s for Stroop, q_{si} for Simon, and q_{fl} for flanker conflicts). Data used were: (i) trial congruency o (indicator function; o_s for Stroop, o_{si} for Simon, and o_{fl} for flanker congruency), assumed to be generated by the conflict probability; (ii) reaction time RT, assumed to be generated by a drift-diffusion model with the correct and wrong choice at the two bounds and the midpoint as the starting point. The statistical significance of the drift rates and drift rate bias coefficients was determined directly by comparing the posterior distribution of the group-level parameters to each other or to zero.

Neurons were selected using a Spearman's rank correlation between spike counts and a given performance monitoring variable ($P < 0.05$; error versus correct, conflict level, and conflict probability). Epochs used for selection were the baseline period (from 1.5 s before stimulus onset to stimulus onset), the ex ante epoch (a 500-ms window centered on the midpoint between stimulus onset and button press), and the ex post period (1-s window start at button press).

For decoding, one trial for each neuron from one condition was randomly selected and concatenated (across neurons) to form a testing data matrix. The rest of the trials were averaged for each condition and concatenated to form a training data matrix ("training means"). For within-task analysis, this matrix was used to compute the coding dimensions (by subtraction between conditions). For cross-task generalization analyses, training data matrices from both tasks were concatenated, and demixed PCA (*57*) was used to extract the task-invariant coding dimensions (except in fig. S12, in which cross-task decoding was evaluated by training in each task separately). Both testing and training data were then projected onto the identified coding dimensions. The predicted labels for the testing data were assigned according to the label of the nearest neighbor of the training data. This decoding procedure was repeated 1000 times (resulting in 1000 single-trial testing data matrices and the corresponding training data matrices). To determine statistical significance, we permuted the trial labels 500 times, and for each permutation, we repeated all above steps to generate a null distribution.

To assess the compositionality of conflict representation by decoding (Fig. 4D), the following coding dimensions were used: one flanked by nonconflict and si training means (Fig. 4A, blue), one flanked by fl and sf training means (blue), one flanked by fl and nonconflict trial averages (orange), and one flanked by si and sf training means (orange). Held-out testing data from conditions flanking one of the blue (or orange) pairs of edges were then projected to the opposite edge in the pair and classified by the training data defining this edge.

REFERENCES AND NOTES

1. E. Koechlin, C. Ody, F. Kouneiher, The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, 1181–1185 (2003). doi: 10.1126/science.1088545; pmid: 14615530
2. D. Badre, M. D'Esposito, Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *J. Cogn. Neurosci.* **19**, 2082–2099 (2007). doi: 10.1162/jocn.2007.19.12.2082; pmid: 17892391
3. D. Badre, D. E. Nee, Frontal cortex and the hierarchical control of behavior. *Trends Cogn. Sci.* **22**, 170–188 (2018). doi: 10.1016/j.tics.2017.11.005; pmid: 29229206
4. M. Ullsperger, C. Danielmeier, G. Jocham, Neurophysiology of performance monitoring and adaptive behavior. *Physiol. Rev.* **94**, 35–79 (2014). doi: 10.1152/physrev.00041.2012; pmid: 24382883
5. R. E. Passingham, S. L. Bengtsson, H. C. Lau, Medial frontal cortex: From self-generated action to reflection on one's own performance. *Trends Cogn. Sci.* **14**, 16–21 (2010). doi: 10.1016/j.tics.2009.11.001; pmid: 19969501
6. J. D. Cosman, K. A. Lowe, W. Zinke, G. F. Woodman, J. D. Schall, Prefrontal control of visual distraction. *Curr. Biol.* **28**, 414–420.e3 (2018). doi: 10.1016/j.cub.2017.12.023; pmid: 29358071
7. C. Danielmeier, T. Eichele, B. U. Forstmann, M. Tittgemeyer, M. Ullsperger, Posterior medial frontal cortex activity predicts post-error adaptations in task-related visual and motor areas. *J. Neurosci.* **31**, 1780–1789 (2011). doi: 10.1523/JNEUROSCI.4299-10.2011; pmid: 21289188
8. T. Egner, J. Hirsch, Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nat. Neurosci.* **8**, 1784–1790 (2005). doi: 10.1038/nn1594; pmid: 16286928
9. B. A. Purcell, R. Kiani, Neural mechanisms of post-error adjustments of decision policy in parietal cortex. *Neuron* **89**, 658–671 (2016). doi: 10.1016/j.neuron.2015.12.027; pmid: 26804992
10. J. G. Kerns *et al*., Anterior cingulate conflict monitoring and adjustments in control. *Science* **303**, 1023–1026 (2004). doi: 10.1126/science.1089910; pmid: 14963333
11. A. W. MacDonald 3rd, J. D. Cohen, V. A. Stenger, C. S. Carter, Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science* **288**, 1835–1838 (2000). doi: 10.1126/science.288.5472.1835; pmid: 10846167
12. E. K. Miller, J. D. Cohen, An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001). doi: 10.1146/annurev.neuro.24.1.167; pmid: 11283309
13. A. Shenhav, M. M. Botvinick, J. D. Cohen, The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron* **79**, 217–240 (2013). doi: 10.1016/j.neuron.2013.07.007; pmid: 23889930
14. R. Nigbur, M. X. Cohen, K. R. Ridderinkhof, B. Stürmer, Theta dynamics reveal domain-specific control over stimulus and response conflict. *J. Cogn. Neurosci.* **24**, 1264–1274 (2012). doi: 10.1162/jocn_a_00128; pmid: 21861681
15. S. D. McDougle *et al*., Credit assignment in movement-dependent reinforcement learning. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 6797–6802 (2016). doi: 10.1073/pnas.1523669113; pmid: 27247404
16. M. Sarafyazd, M. Jazayeri, Hierarchical reasoning by neural circuits in the frontal cortex. *Science* **364**, eaav8911 (2019). doi: 10.1126/science.aav8911; pmid: 31097640
17. D. Badre, A. D. Wagner, Selection, integration, and conflict monitoring: Assessing the nature and generality of prefrontal cognitive control mechanisms. *Neuron* **41**, 473–487 (2004). doi: 10.1016/S0896-6273(03)00851-1; pmid: 14766185
18. A. Genovesio, P. J. Brasted, A. R. Mitz, S. P. Wise, Prefrontal cortex activity related to abstract response strategies. *Neuron* **47**, 307–320 (2005). doi: 10.1016/j.neuron.2005.06.006; pmid: 16039571
19. P. Domenech, S. Rheims, E. Koechlin, Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex. *Science* **369**, eabb0184 (2020). doi: 10.1126/science.abb0184; pmid: 32855307
20. J. R. Wessel, A. R. Aron, On the globality of motor suppression: Unexpected events and their influence on behavior and cognition. *Neuron* **93**, 259–280 (2017). doi: 10.1016/j.neuron.2016.12.013; pmid: 28103476
21. Y. Niv, N. D. Daw, D. Joel, P. Dayan, Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology* **191**, 507–520 (2007). doi: 10.1007/s00213-006-0502-4; pmid: 17031711
22. J. F. Cavanagh *et al*., Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nat. Neurosci.* **14**, 1462–1467 (2011). doi: 10.1038/nn.2925; pmid: 21946325
23. E. A. Crone, R. J. M. Somsen, B. Van Beek, M. W. Van Der Molen, Heart rate and skin conductance analysis of antecendents

Downloaded from https://www.science.org at California Institute of Technology on May 05, 2022

and consequences of decision making. *Psychophysiology* **41**, 531–540 (2004). doi: 10.1111/j.1469-8986.2004.00197.x; pmid: 15189476
24. R. B. Ebitz, M. L. Platt, Neuronal activity in primate dorsal anterior cingulate cortex signals task conflict and predicts adjustments in pupil-linked arousal. *Neuron* **85**, 628–640 (2015). doi: 10.1016/j.neuron.2014.12.053; pmid: 25654259
25. P. R. Murphy, E. Boonstra, S. Nieuwenhuis, Global gain modulation generates time-dependent urgency during perceptual choice in humans. *Nat. Commun.* **7**, 13526 (2016). doi: 10.1038/ncomms13526; pmid: 27882927
26. E. Eldar, R. B. Rutledge, R. J. Dolan, Y. Niv, Mood as representation of momentum. *Trends Cogn. Sci.* **20**, 15–24 (2016). doi: 10.1016/j.tics.2015.07.010; pmid: 26545853
27. R. B. Rutledge, N. Skandali, P. Dayan, R. J. Dolan, A computational and neural model of momentary subjective well-being. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 12252–12257 (2014). doi: 10.1073/pnas.1407535111; pmid: 25092308
28. A. J. Shackman *et al*., The integration of negative affect, pain and cognitive control in the cingulate cortex. *Nat. Rev. Neurosci.* **12**, 154–167 (2011). doi: 10.1038/nrn2994; pmid: 21331082
29. T. Paus, Primate anterior cingulate cortex: Where motor control, drive and cognition interface. *Nat. Rev. Neurosci.* **2**, 417–424 (2001). doi: 10.1038/35077500; pmid: 11389475
30. P. Nachev, C. Kennard, M. Husain, Functional role of the supplementary and pre-supplementary motor areas. *Nat. Rev. Neurosci.* **9**, 856–869 (2008). doi: 10.1038/nrn2478; pmid: 18843271
31. F. Bonini *et al*., Action monitoring and medial frontal cortex: Leading role of supplementary motor area. *Science* **343**, 888–891 (2014). doi: 10.1126/science.1247412; pmid: 24558161
32. C. S. Carter *et al*., Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science* **280**, 747–749 (1998). doi: 10.1126/science.280.5364.747; pmid: 9563953
33. Z. Fu *et al*., Single-neuron correlates of error monitoring and post-error adjustments in human medial frontal cortex. *Neuron* **101**, 165–177.e5 (2019). doi: 10.1016/j.neuron.2018.11.016; pmid: 30528064
34. T. Gazit *et al*., The role of mPFC and MTL neurons in human choice under goal-conflict. *Nat. Commun.* **11**, 3192 (2020). doi: 10.1038/s41467-020-16908-z; pmid: 32581214
35. S. R. Heilbronner, B. Y. Hayden, Dorsal anterior cingulate cortex: A bottom-up view. *Annu. Rev. Neurosci.* **39**, 149–170 (2016). doi: 10.1146/annurev-neuro-070815-013952; pmid: 27090954
36. N. Kolling *et al*., Value, search, persistence and model updating in anterior cingulate cortex. *Nat. Neurosci.* **19**, 1280–1285 (2016). doi: 10.1038/nn.4382; pmid: 27669988
37. A. Sajad, D. C. Godlove, J. D. Schall, Cortical microcircuitry of performance monitoring. *Nat. Neurosci.* **22**, 265–274 (2019). doi: 10.1038/s41593-018-0309-8; pmid: 30643297
38. S. A. Sheth *et al*., Human dorsal anterior cingulate cortex neurons mediate ongoing behavioural adaptation. *Nature* **488**, 218–221 (2012). doi: 10.1038/nature11239; pmid: 22722841
39. H. Tang *et al*., Cascade of neural processing orchestrates cognitive control in human frontal cortex. *eLife* **5**, e12352 (2016). doi: 10.7554/eLife.12352; pmid: 26888070
40. S. Ito, V. Stuphorn, J. W. Brown, J. D. Schall, Performance monitoring by the anterior cingulate cortex during saccade countermanding. *Science* **302**, 120–122 (2003). doi: 10.1126/science.1087847; pmid: 14526085
41. V. Stuphorn, T. L. Taylor, J. D. Schall, Performance monitoring by the supplementary eye field. *Nature* **408**, 857–860 (2000). doi: 10.1038/35048576; pmid: 11130724
42. C. Wang, I. Ulbert, D. L. Schomer, K. Marinkovic, E. Halgren, Responses of human anterior cingulate cortex microdomains to error detection, conflict monitoring, stimulus-response mapping, familiarity, and orienting. *J. Neurosci.* **25**, 604–613 (2005). doi: 10.1523/JNEUROSCI.4151-04.2005; pmid: 15659596
43. Z. M. Williams, G. Bush, S. L. Rauch, G. R. Cosgrove, E. N. Eskandar, Human anterior cingulate neurons and the integration of monetary reward with motor responses. *Nat. Neurosci.* **7**, 1370–1375 (2004). doi: 10.1038/nn1354; pmid: 15558064
44. S. Bernardi *et al*., The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967.e21 (2020). doi: 10.1016/j.cell.2020.09.031; pmid: 33058757
45. C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, K. D. Harris, High-dimensional geometry of population responses in visual cortex. *Nature* **571**, 361–365 (2019). doi: 10.1038/s41586-019-1346-5; pmid: 31243367
46. J. J. DiCarlo, D. D. Cox, Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007). doi: 10.1016/j.tics.2007.06.010; pmid: 17631409
47. S. Fusi, E. K. Miller, M. Rigotti, Why neurons mix: High dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016). doi: 10.1016/j.conb.2016.01.010; pmid: 26851755
48. M. Rigotti *et al*., The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013). doi: 10.1038/nature12160; pmid: 23685452
49. G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, X.-J. Wang, Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019). doi: 10.1038/s41593-018-0310-2; pmid: 30643294
50. J. Minxha, R. Adolphs, S. Fusi, A. N. Mamelak, U. Rutishauser, Flexible recruitment of memory-based choice representations by the human medial frontal cortex. *Science* **368**, eaba3313 (2020). doi: 10.1126/science.aba3313; pmid: 32586990
51. T. S. Braver, The variable nature of cognitive control: A dual mechanisms framework. *Trends Cogn. Sci.* **16**, 106–113 (2012). doi: 10.1016/j.tics.2011.12.010; pmid: 22245618
52. J. Jiang, J. Beck, K. Heller, T. Egner, An insula-frontostriatal network mediates flexible cognitive control by adaptively predicting changing control demands. *Nat. Commun.* **6**, 8165 (2015). doi: 10.1038/ncomms9165; pmid: 26391305
53. J. Jiang, K. Heller, T. Egner, Bayesian modeling of flexible cognitive control. *Neurosci. Biobehav. Rev.* **46**, 30–43 (2014). doi: 10.1016/j.neubiorev.2014.06.001; pmid: 24929218
54. T. E. J. Behrens, M. W. Woolrich, M. E. Walton, M. F. S. Rushworth, Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007). doi: 10.1038/nn1954; pmid: 17676057
55. Four additional Bayesian models were considered for model comparison: (i) DDM only, with no CP estimation; (ii) DDM with drift rate biases depending on the posterior means of CP (inferred separately); (iii) DDM with drift rate biases depending on CP updated by a deterministic learning rule with a single learning rate; and (iv) DDM with drift rate biases depending on previous trial conflict. The full conflict estimation model outperformed these alternative models (see table S2 for measures of model comparison and table S3 for model weights). Comparison with the first model suggests that subjects used a flexible instead of fixed learning rate in estimating CP. Comparison with the second model suggests that not only the CP but also its uncertainty provides useful information for predicting RT. The comparison with the fourth model, consistent with prior work (*52*), demonstrates that participants incorporated conflict from multiple past trials.
56. G. Gratton, M. G. Coles, E. Donchin, Optimizing the use of information: Strategic control of activation of responses. *J. Exp. Psychol. Gen.* **121**, 480–506 (1992). doi: 10.1037/0096-3445.121.4.480; pmid: 1431740
57. D. Kobak *et al*., Demixed principal component analysis of neural population data. *eLife* **5**, e10989 (2016). doi: 10.7554/eLife.10989; pmid: 27067378
58. J. Jiang, T. Egner, Using neural pattern classifiers to quantify the modularity of conflict-control mechanisms in the human brain. *Cereb. Cortex* **24**, 1793–1805 (2014). doi: 10.1093/cercor/bht029; pmid: 23402762
59. T. A. Niendam *et al*., Meta-analytic evidence for a superordinate cognitive control network subserving diverse executive functions. *Cogn. Affect. Behav. Neurosci.* **12**, 241–268 (2012). doi: 10.3758/s13415-011-0083-5; pmid: 22282036
60. P. A. Kragel *et al*., Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex. *Nat. Neurosci.* **21**, 283–289 (2018). doi: 10.1038/s41593-017-0051-7; pmid: 29292378
61. S. Tsujimoto, A. Genovesio, S. P. Wise, Frontal pole cortex: Encoding ends at the end of the endbrain. *Trends Cogn. Sci.* **15**, 169–176 (2011). doi: 10.1016/j.tics.2011.02.001; pmid: 21388858
62. F. A. Mansouri, E. Koechlin, M. G. P. Rosa, M. J. Buckley, Managing competing goals—a key role for the frontopolar cortex. *Nat. Rev. Neurosci.* **18**, 645–657 (2017). doi: 10.1038/nrn.2017.111; pmid: 28951610
63. V. Mante, D. Sussillo, K. V. Shenoy, W. T. Newsome, Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013). doi: 10.1038/nature12742; pmid: 24201281
64. N. Yeung, C. Summerfield, Metacognition in human decision-making: Confidence and error monitoring. *Philos. Trans. R. Soc. London B Biol. Sci.* **367**, 1310–1321 (2012). doi: 10.1098/rstb.2011.0416; pmid: 22492749
65. S. M. Fleming, R. J. Dolan, The neural basis of metacognitive ability. *Philos. Trans. R. Soc. London Ser. B* **367**, 1338–1349 (2012). doi: 10.1098/rstb.2011.0417; pmid: 22492751
66. J. Morales, H. Lau, S. M. Fleming, Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *J. Neurosci.* **38**, 3534–3546 (2018). doi: 10.1523/JNEUROSCI.2360-17.2018; pmid: 29519851
67. N. U. F. Dosenbach *et al*., A core system for the implementation of task sets. *Neuron* **50**, 799–812 (2006). doi: 10.1016/j.neuron.2006.04.031; pmid: 16731517
68. M. F. S. Rushworth, M. E. Walton, S. W. Kennerley, D. M. Bannerman, Action sets and decisions in the medial frontal cortex. *Trends Cogn. Sci.* **8**, 410–417 (2004). doi: 10.1016/j.tics.2004.07.009; pmid: 15350242
69. M. W. Cole *et al*., Multi-task connectivity reveals flexible hubs for adaptive task control. *Nat. Neurosci.* **16**, 1348–1355 (2013). doi: 10.1038/nn.3470; pmid: 23892552
70. M. M. Botvinick, T. S. Braver, D. M. Barch, C. S. Carter, J. D. Cohen, Conflict monitoring and cognitive control. *Psychol. Rev.* **108**, 624–652 (2001). doi: 10.1037/0033-295X.108.3.624; pmid: 11488380
71. N. Yeung, M. M. Botvinick, J. D. Cohen, The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychol. Rev.* **111**, 931–959 (2004). doi: 10.1037/0033-295X.111.4.931; pmid: 15482068
72. C.-C. Lo, X.-J. Wang, Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nat. Neurosci.* **9**, 956–963 (2006). doi: 10.1038/nn1722; pmid: 16767089
73. D. Subramanian, A. Alers, M. A. Sommer, Corollary discharge for action and cognition. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **4**, 782–790 (2019). doi: 10.1016/j.bpsc.2019.05.010; pmid: 31351985
74. W. H. Alexander, J. W. Brown, Medial prefrontal cortex as an action-outcome predictor. *Nat. Neurosci.* **14**, 1338–1344 (2011). doi: 10.1038/nn.2921; pmid: 21926982
75. P. Haggard, Sense of agency in the human brain. *Nat. Rev. Neurosci.* **18**, 196–207 (2017). doi: 10.1038/nrn.2017.14; pmid: 28251993
76. T. Kawai, H. Yamada, N. Sato, M. Takada, M. Matsumoto, 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 (2019). doi: 10.1093/cercor/bhy103; pmid: 29722795
77. N. Brunel, X.-J. Wang, Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *J. Comput. Neurosci.* **11**, 63–85 (2001). doi: 10.1023/A:1011204814320; pmid: 11524578
78. Z. Fu *et al*., Data for: The geometry of domain-general performance monitoring in the human medial frontal cortex, Open Science Framework (2022); https://osf.io/42r9c/.

ACKNOWLEDGMENTS

We thank the members of the Adolphs and Rutishauser labs, L. J. Jin, and S. Dong 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 (R.A.). **Author contributions:** Z.F., R.A., and U.R. designed the study. Z.F. and U.R. collected and analyzed the data and implemented analysis procedures. Z.F., U.R., A.N.M., and R.A. wrote the paper. D.B. acquired and analyzed the behavioral control data. J.M.C. and C.M.R. provided patient care and facilitated experiments. A.N.M. performed surgery. **Competing interest:** The authors have no competing interests to declare. **Data and materials availability:** Data needed to reproduce results have been deposited at Open Science Framework (OSF) (*78*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abm9922
Materials and Methods
Figs. S1 to S15
Tables S1 to S3
References (*79*–*98*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

Submitted 27 October 2021; accepted 1 April 2022
10.1126/science.abm9922

Downloaded from https://www.science.org at California Institute of Technology on May 05, 2022

Science

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

Zhongzheng FuDanielle BeamJeffrey M. ChungChrystal M. ReedAdam N. MamelakRalph AdolphsUeli Rutishauser

Science, 376 (6593), eabm9922. • DOI: 10.1126/science.abm9922

The neural basis of cognitive control

What are the neural mechanisms that enable humans to flexibly control and monitor their own actions in various non-routine and novel tasks? Fu *et al*. recorded the activity of more than 1000 neurons in the medial frontal cortex of human epilepsy patients while they performed complex cognitive tasks. They found that domain-general and domain-specific performance monitoring neurons were intermixed within this brain region. The population activity gave rise to a geometry that allowed domain-general signals to be read out with more than 90% accuracy on single trials while at the same time retaining the ability to separate different conflict conditions. These results show how the human medial frontal cortex resolves the fundamental trade-off between task generalization and specialization, which is critical for cognitive flexibility. —PRS

View the article online
https://www.science.org/doi/10.1126/science.abm9922
Permissions
https://www.science.org/help/reprints-and-permissions

Downloaded from https://www.science.org at California Institute of Technology on May 05, 2022

Use of this article is subject to the Terms of service

Science (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.
Copyright © 2022 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works