Task interference as a neuronal basis for the cost of cognitive flexibility

Authors:

Cheng Xue*[1], Sol K. Markman*[1,2], Ruoyi Chen[3], Lily E. Kramer[1], Marlene R. Cohen[1]

[1] Department of Neurobiology, University of Chicago, IL, USA
[2] Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology, MA, USA
[3] Department of Biological Sciences, Carnegie Mellon University, PA, USA

Disclosures: None

Correspondence can be addressed to Cheng Xue (cxue@uchicago.edu) and Marlene R. Cohen (lead contact, marlenecohen@uchicago.edu)

Abstract:

Humans and animals have an impressive ability to juggle multiple tasks in a constantly changing environment. This flexibility, however, leads to decreased performance under uncertain task conditions. Here, we combined monkey electrophysiology, human psychophysics, and artificial neural network modeling to investigate the neuronal mechanisms of this performance cost. We developed a behavioural paradigm to measure and influence participants' decision-making and perception in two distinct perceptual tasks. Our data revealed that both humans and monkeys, unlike an artificial neural network trained for the same tasks, make less accurate perceptual decisions when the task is uncertain. We generated a mechanistic hypothesis by comparing this neural network trained to produce correct choices with another network trained to replicate the participants' choices. We hypothesized, and confirmed with further behavioural, physiological, and causal experiments, that the cost of task flexibility comes from what we term task interference. Under uncertain conditions, interference between different tasks causes errors because it results in a stronger representation of irrelevant task features and entangled neuronal representations of different features. Our results suggest a tantalizing, general hypothesis: that cognitive capacity limitations, both in health and disease, stem from interference between neural representations of different stimuli, tasks, or memories.

Introduction:

In the face of the inherent uncertainty and unexpected changes in natural environments, humans and animals have evolved remarkable cognitive flexibility, enabling them to seamlessly switch between tasks as the situation demands. However, this cognitive flexibility comes with a cost: when we are uncertain which task should be performed (typically during and shortly before or after task switching), we take longer to perform the task we choose and do so with lower accuracy (1-4).

Numerous non-exclusive hypotheses have been proposed based on behavioural evidence to explain the cost of task flexibility. These include limits on processing capacity (4), excessive attention to irrelevant stimuli or features (5-8) and proactive interference between competing tasks (1, 2, 9-12). These phenomenological explanations of behaviour do not directly address the neural mechanisms that underly task switching costs and related limits on other forms of cognition.

Here, we propose a novel neural mechanism that integrates and elucidates the roles and connections among various hypotheses addressing cognitive flexibility cost. We used a multidisciplinary approach including behavioural experiments in humans and rhesus monkeys, multi-neuron, multi-area recordings and causal manipulations in monkeys, and artificial neural network modeling to generate and test neural hypotheses. The core of our approach is our two-feature visual discrimination task, in which participants must discriminate one visual feature and ignore the other, inferring the implicitly changing task rule that determines which feature is relevant, based on their history of stimuli, choices and rewards.

Our findings suggest that the cost of task flexibility arises from stronger task interference under uncertain conditions. Such interference is a joint product of retaining task-irrelevant visual information and the entanglement of the neural representations of unrelated features. This phenomenon likely underlies the cost of cognitive flexibility across species, systems, and health states, offering promising avenues for advancing our understanding and addressing cognitive limitations in both basic science and translational research.

Results:

Task switching can have behavioural costs.

We designed a two-feature discrimination task that allows us to measure and manipulate perception and task certainty, pushing participants to flexible (task uncertain) and engaged (task certain) cognitive states(3) (Figure 1A). We trained 220 online human participants and two rhesus monkeys (*Macaca mulatta*; both male, 12 and 9 kg) to discriminate changes in two visual features of subsequently presented Gabor stimuli (spatial frequency and orientation for human participants; spatial frequency and location for monkey participants). Both features changed on each trial near perceptual threshold (average perceptual performance is 79% for monkeys, and 82% for humans), but only one was relevant. Participants indicated both their chosen task and perceptual discrimination on every trial, and only correct discriminations of the relevant feature were rewarded (Figure 1A-B). The relevant feature switched without a cue with a constant low probability on each trial (2.5% for monkeys and 10% for humans).

The task structure meant that the participants' certainty about which task they would be rewarded for performing fluctuated throughout the session based on their history of stimuli, choices, and rewards. A rewarded trial suggested a high likelihood of the same task on the next trial (97.5% for monkeys, 90% for humans), thus leaving participants with higher task certainty. An unrewarded trial, however, led to task uncertainty because participants received no explicit feedback about the error type (i.e., whether the task had switched, they had made a perceptual

error, or both). For simplicity, we will refer to trials as having high or low task certainty based on whether they were immediately preceded by a rewarded trial or an unrewarded trial. This dichotomy proves to be an effective separation between high and low task certainty, supported by task-related neuronal representations in the parietal cortex, and distinguished behavioural patterns not solely attributable to recent rewards or trial position in a certain block (3). In the following results, we will compare the trial condition under low task certainty (following an unrewarded trial) vs. high task certainty (following a rewarded trial) to evaluate the cost for flexibility.

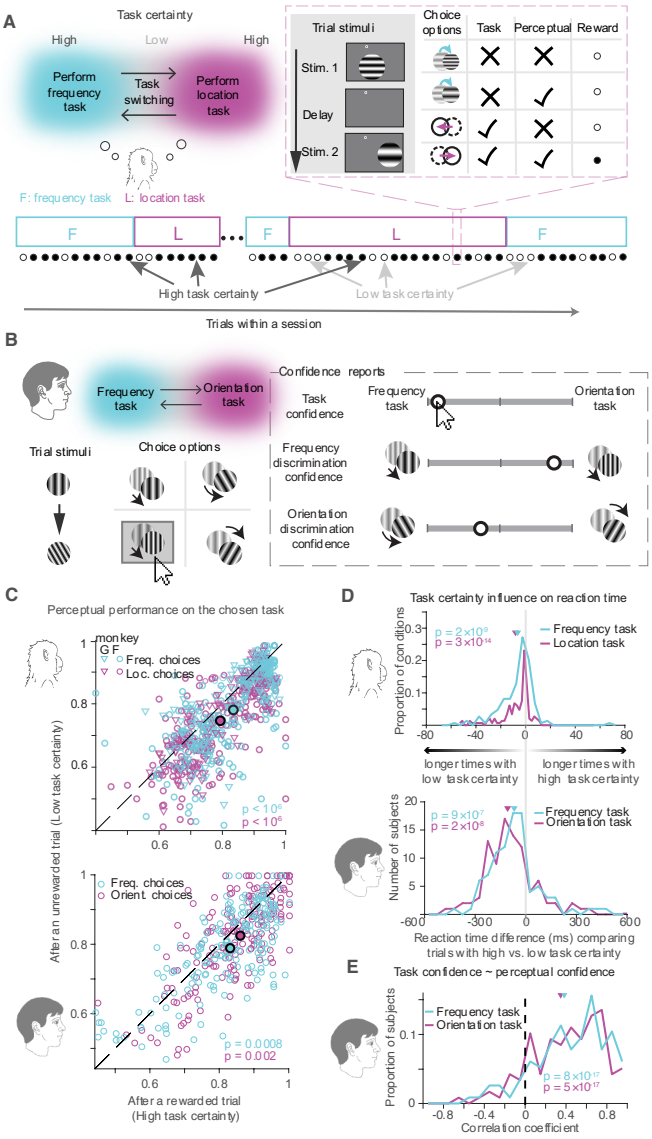

FIG 1: Tasks and behavioural cost of flexibility. (A) Monkeys discriminated either the spatial frequency (cyan) or location (magenta) of two sequentially presented Gabor stimuli while ignoring changes in the irrelevant feature. The relevant feature is not explicitly cued. There was a fixed, low probability of a task switch after each trial, leading to task blocks of variable

length (cyan and magenta rectangles). Participants had to infer the relevant feature from their history of choices and feedback. Participants were rewarded only if they made the correct task choice and perceptual judgment (they indicated both on each trial by choosing one of four possible targets that represent the choices schematized in the first column of the table in A). Therefore, task certainty was generally high following rewarded trials and low following unrewarded trials. (B) Human participants performed the same task as the monkeys, alternating between orientation discrimination and frequency discrimination. On a small proportion of trials, some human participants were asked to report their confidence about their task choice and perceptual discrimination of both features on a continuous scale. (C) Behavioural cost of flexibility in perceptual performance for monkeys (top) and humans (bottom). Each open symbol represents perceptual performance for a set of trials with identical stimuli. Under low task certainty (following an unrewarded trial), both species show worse perceptual performance under low task certainty (following an unrewarded trial; ordinates) than under high task certainty (following a rewarded trial; abscissae; $p=0.002$ for human orientation choices, $p=0.0008$ for human frequency choices; $p<10^{-6}$ for monkey location and frequency choices. Filled symbols denote mean performance. (D) Behavioural cost of flexibility in reaction time. The upper (monkey) and lower (human) histograms show distributions of the differences between reaction times under high and low task certainty for matched stimuli. The histograms are shifted to the left, indicating that participants reported their perceptual discriminations more slowly under low task certainty than under high task certainty. $p<10^{6}$ for both species and both tasks. (E) Behavioural cost of flexibility in confidence. Cyan and magenta curves show the distribution of correlation coefficients between task confidence and perceptual confidence on the chosen task for human participants. $p<10^{6}$ for both tasks.

Both humans and monkeys demonstrate a cost of flexibility. Our participants perform the perceptual task better (Figure 1C), faster (Figure 1D), and with greater confidence (Figure 1E) following a rewarded trial, when task certainty is high.

Recurrent neural networks trained on behavioural data can generate mechanistic hypotheses

To generate hypotheses about the neural mechanisms underlying this cost of flexibility, we trained recurrent neural networks (RNNs) to perform the same two-feature discrimination task and then analyzed the internal dynamics of the models. We used two training strategies to create two distinct, comparable models. Both models received as inputs information about features of the stimuli on the current trial and the history of stimuli, choices, and rewards over the past nine trials (see Methods: RNN models and analyses). The first model, the “correct choice network,” was trained to produce the correct choice given those inputs (Figure 2A; see Methods). The correct choice network did not exhibit a cost of flexibility at any point during training (Figure 2B), which indicates that the cost of flexibility is not somehow related to the structure or statistics of our task; nor does it necessarily appear during the intermediate stage of task learning. The second model, the “monkey choice network,” was trained to predict the monkeys’ actual choices

(Figure 2A). Because of the way it was trained, unlike the correct choice network, the monkey choice network exhibited the cost of flexibility characteristic of the behaviour of our monkeys and human participants, which persisted from about halfway through training (Figure 2B and 2C). To model reaction times, we applied a collapsing decision boundary to the four choice output units such that the reaction time is given by the first timepoint in which any choice unit's activity crosses the boundary. The boundary parameters were chosen such that the mean reaction times of the monkey choice network after rewarded and unrewarded trials were comparable to those of the monkeys. We compared these reaction times with those of the correct choice network using the same decision boundary (Figure 2D). The correct choice network was faster than the monkey choice network both after rewarded trials and unrewarded trials. Notably, this difference was greater after unrewarded trials, indicating that task certainty has a greater effect on the time course of choice output activity in the monkey choice network.

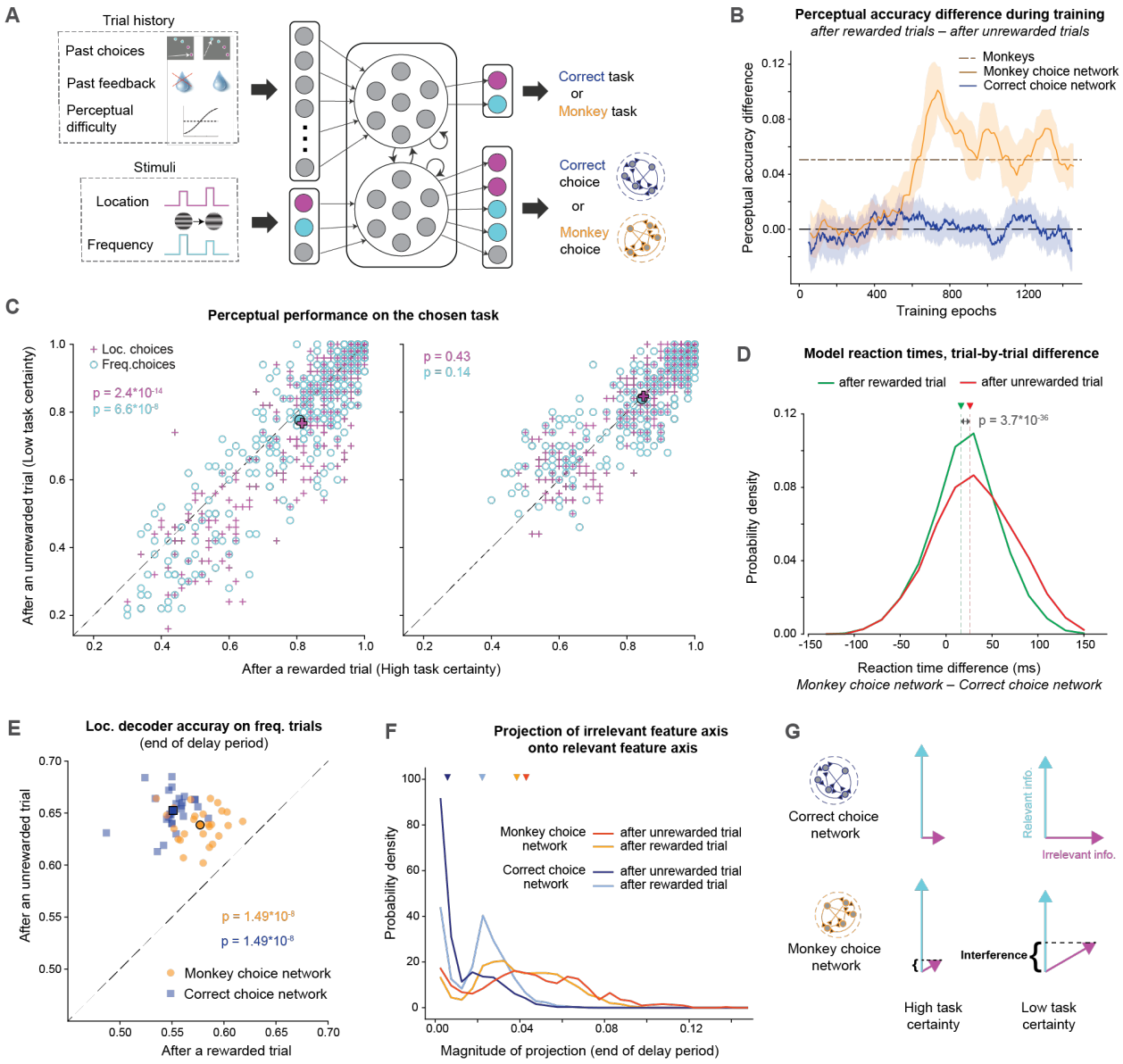

FIG 2: Generating mechanistic hypotheses using recurrent neural networks. (A) Model schematic for RNNs designed to either predict the correct choice (correct choice network) or predict the monkeys' choice on each trial (monkey choice network). (B) The difference in perceptual accuracy (after rewarded trials – after unrewarded trials) throughout training for both models. The two monkeys' difference in perceptual accuracy was on average about 5%, as indicated by the brown dashed line. The monkey choice network approaches this difference, while the correct choice network remains around zero (shaded regions show 95% confidence intervals from sliding window averaging). (C) The monkey choice network (left) displays a cost of flexibility after training, as shown by worse perceptual performance after an unrewarded, low task certainty trial (location choices: $p=2.4\times10^{-14}$, frequency choices: $p=6.6\times10^{-8}$). In contrast, the correct choice network (right) did not show a significant cost of flexibility (location choices: p=0.43, frequency choices: p=0.14). Each point corresponds to a set of accuracies calculated from 50 trials with identical feature change amounts. (D) The monkey choice network displays a larger effect of task uncertainty on reaction time compared to the correct choice network. The plot depicts histograms of reaction times when the same collapsing decision boundary is applied to the outputs of both RNNs (see Methods). Overall, the monkey choice network has longer reaction times than the correct choice network given the same inputs (after unrewarded trials: $p=1.7\times10^{-271}$, after rewarded trials: $p<10^{-300}$). The difference between models is greater after unrewarded than after rewarded trials ($p=3.7\times10^{-36}$, Wilcoxon rank sum test. (E) At the end of the delay period between stimulus presentations (400 ms after first stimulus presentation), information about the irrelevant feature (e.g. location on trials when the model chose the spatial frequency task) is encoded better following non-rewarded than rewarded trials (points are above the diagonal for both models; $p=1.49\times10^{-8}$ for both). Each point represents a set of trials with an identical first stimulus and shows the performance of a corresponding cross-validated linear decoder. (F) The two features are represented more orthogonally in the correct choice network than in the monkey choice network at the end of the delay period (400 ms after first stimulus presentation). Histograms of the magnitude of the projection of the irrelevant feature axis onto the relevant feature axis for the monkey choice and correct choice networks after unrewarded and rewarded trials. The unit axis for each feature was computed using distance covariance analysis and scaled by the corresponding distance covariance (see Methods for details). All four distributions are significantly different (Wilcoxon rank sum tests: monkey choice network after non-reward > correct choice network after non-reward: $p=6.1\times10^{-106}$; monkey choice network after reward > correct choice network after reward: $p=3.8\times10^{-142}$; monkey choice network after reward > after non-reward: p=0.001, correct choice network after non-reward > after reward: $p=1.8\times10^{-35}$). Triangle markers represent the medians. (G) A schematic of how overrepresentation of the irrelevant feature when the task is uncertain (depicted in E) combined with non-orthogonal representations of task variables (depicted in F) leads to interference and thus a cost of flexibility. In an idealized correct choice network (top), the irrelevant feature is more strongly encoded (longer magenta axis) under low task certainty, however, it varies orthogonally to the relevant feature, thus avoiding interference.

In the monkey choice network (bottom), the feature axes are no longer orthogonal, leading to more interference under low task certainty.

Task uncertainty leads to task interference

The observation that our monkey choice network, but not our correct choice network, demonstrates costs of flexibility gives us a platform for generating hypotheses about the mechanistic origins of this cost. We investigated the dynamics of the two networks and identified two features that could in principle underlie the cost of flexibility. First, consistent with the concepts of attentional and task-set inertia (1, 2, 5-7, 9, 10, 12), both models represented the irrelevant feature better under low task certainty than under high task certainty (orange data points, Figure 2E). When the task was more certain (following rewarded trials), the relevant feature was encoded throughout the entire delay period, while information about the irrelevant feature decayed nearly completely (Figure S2). This certainty-dependent 'forgetting' of irrelevant information occurred even in the correct choice network (blue data points, Figure 2E), which did not show a cost of flexibility, so it is on its own not a sufficient explanation of the cost of flexibility in biological organisms.

The second feature of the models is not predicted by common accounts of task switch costs. In the monkey choice network, the representations of the two visual features of the first stimulus become non-orthogonal before the onset of the second stimulus presentation (Figures 2F, S3). This tangling of the representations causes the monkey choice network to mix up information about the two features (e.g. judgments of spatial frequency are influenced by the location of the stimulus), which reduces accuracy. We can visualize the problem by identifying the axes that best represent information about each visual feature in a space in which the activity of each unit in the hidden layer is one dimension. These axes are the directions in population space along which there is most variation in responses to different values of that feature. If information about the two features is encoded independently, the axes best representing them will be orthogonal. However, in the monkey choice network, those axes rotate such that by the end of the delay period they are no longer orthogonal (Figures 2F, S3(B)). This non-orthogonality means that information about the two features is no longer independent, leading to 'switch errors' that worsen perceptual performance. In contrast, the correct choice network maintains nearly orthogonal feature axes, especially after unrewarded trials, such that their dot product remains closer to zero (Figures 2F, S3).

We hypothesize that the cost of flexibility is the joint product of these two mechanisms (Figure 2G). The increased representation of irrelevant information when the task is uncertain is not a problem when information about the two features is encoded independently (as in the correct choice network). The independent representations make it straightforward to ignore irrelevant information, even when it is robustly represented. However, when information about the two features is not independent, as in the monkey choice network, the irrelevant information is confounded with information about the relevant feature, worsening perceptual performance.

We tested predictions of this task interference hypothesis using new behavioural, neuronal, and causal experiments designed to identify the types of errors associated with task uncertainty.

Behavioural evidence for task interference

We first looked for behavioural evidence supporting the first part of our task interference hypothesis, the information deemed irrelevant for the chosen task is more strongly retained under low task certainty. On a small subset of trials, we asked human participants to report one of the features (orientation or frequency) of the first stimulus after a delay (Figure 3A). Consistent with the idea that irrelevant information is maintained longer in short-term memory when the task is uncertain, participants had higher accuracy (smaller errors) when estimating the believed-irrelevant, but not the relevant, feature on low task certainty trials (Figure 3B).

Next, we looked for behavioural evidence supporting the hypothesis that the two features are entangled when the task is uncertain. The monkey choice network predicted that under uncertain task conditions, participants' choices about one feature (e.g., spatial frequency) are influenced by the other feature (e.g., location). Consistent with this hypothesis, the choices of humans (Figure 3C upper panel) and monkeys (Figure 3C lower panel) were more strongly related to the irrelevant stimulus feature following a non-reward (when the task was relatively uncertain) than a reward (when the task was more certain).

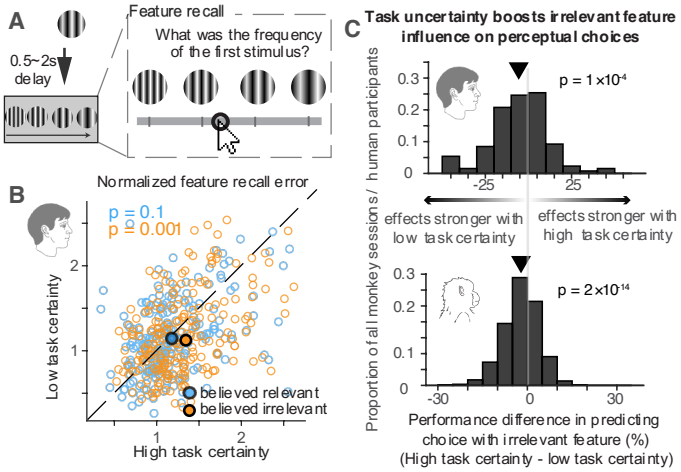

FIG 3: Behavioural evidence for the task interference hypothesis. (A) In a subset of the trials, we asked human participants to reproduce one of the features (spatial frequency or orientation) of the first stimulus according to their memory following a delay. The plot depicts the experimental design, including the slider used to reproduce the stimulus. (B) When participants reported that a feature was behaviourally irrelevant, their recall error was better (smaller) for that feature if the preceding trial had been incorrect (low task certainty, y-axis) than correct (high task certainty, x-axis). The orange data points lie significantly below the identity line (p=0.001), indicating that believed-irrelevant information was reported with smaller errors under low task certainty. In contrast, there was no significant relationship between task certainty and participants' ability to remember the believed-relevant feature (p>0.05), although the deviations from the identity line are not significantly different between believed relevant and believed irrelevant features (Wilcoxon rank sum test, p>0.05).

(C) Perceptual choices are significantly more predictable from the irrelevant feature under lower task certainty. For each monkey experiment session and human participant, we predict the perceptual choices on one feature based on the changes in the other feature for trial-number matched conditions of low and high task certainty. Histograms (top: humans, bottom: monkeys) show that across monkey experimental sessions / human participants, the distributions of differences in prediction performances under high and low task certainty were significantly below zero ($p=1\times10^{-4}$ for humans, $p=2\times10^{-14}$ for monkeys), indicating that irrelevant feature is more predictive of perceptual choices under task uncertainty.

Neuronal evidence for task interference

To test the predictions of the feature interference hypothesis on neural representations, we recorded groups of neurons in monkey primary visual cortex (V1), which is known to encode the visual features (frequency and location) that the monkeys discriminated. We analyzed the neurons in a similar way as the units in our model, beginning by using linear regression to identify the axes in V1 population activity space that best encode each visual feature. We found evidence for task interference in both signal and noise. To look for entangled representations in the signal, we tried to decode the frequency value from location encoding axis, and vice versa. Consistent with our hypothesis, under low task certainty, the representations of spatial frequency and location were more entangled, as indicated by better decoder performance under low task certainty (Figure 4A). To investigate noise-based evidence for task interference, we investigated the V1 population responses to the same stimulus across trials. The trial-to-trial fluctuations between projections onto the spatial location and frequency axes were more correlated under low task certainty (Figure 4B).

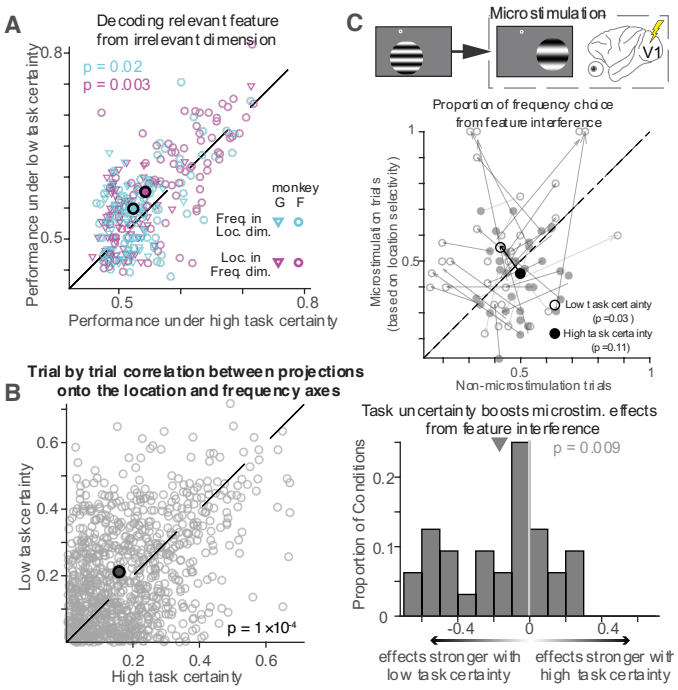

FIG 4: Neuronal and causal evidence for task interference from neural population recordings and electrical microstimulation in monkey V1. (A) Stronger feature entanglement under low task certainty. For each monkey recording session, we identified the dimensions in V1 population space along which we could best linearly decode each visual feature. Using linear decoders, we assessed our ability to decode the relevant feature in the encoding dimension of the irrelevant feature under high (abscissa) and low task certainty (ordinate). Across sessions, for both features, the points are distributed significantly above the identity line ($p=0.003$ for location classifiers in frequency dimension, $p=0.02$ for frequency classifiers in location dimension), indicating increased feature entanglement under low task certainty. (B) Projections of V1 population responses onto the spatial frequency and location axes are more correlated under low task certainty, indicating that the two axes are less orthogonal. The abscissa and ordinate of each data point show the absolute value of the correlation coefficient relating projections onto the two axes for trials with identical stimuli under high and low task certainty. We use the absolute value because there is no a priori prediction about the sign of the correlation, only that nonzero correlation indicates non-orthogonality. The mean of the distribution is above the identity line, indicating that decoding of the two features is more correlated under low task certainty ($p=1\times10^{-4}$). (C) Causal tests of the task interference hypothesis using V1 microstimulation. On a subset of randomly interleaved trials, we microstimulated V1 neurons during the display of the second stimulus. Each point in the scatterplot in the upper panel represents a set of trials with identical stimuli, displaying the proportion of a given spatial frequency choices on trials with microstimulation (ordinate) or without microstimulation (abscissa). If the neuronal population representations of the two features were independent, the points would align along the identity line, while deviation indicates feature interference. The observation that microstimulation has a bigger impact on behaviour under low task certainty is evidence for task interference (Compare microstimulation effects in high certainty: filled grey symbols, not significantly away from identity line, $p=0.11$; vs. low task certainty: open grey symbols, above identity line, $p=0.03$). Black symbols indicate the means of each task certainty condition. The bottom histogram plots the differences between microstimulation effects under low and high task certainty (i.e. how much the arrows in the scatter moved the data points away from the diagonal). The distribution across conditions reveals a significantly negative distribution ($p=0.009$) that confirms stronger feature interference in V1 under low task certainty.

Causal evidence for task interference

Finally, we used a causal manipulation to test the prediction of the task interference hypothesis that the representations of the two features are more entangled under low task certainty. If the representations of the two features are orthogonal, then a causal manipulation that biases judgments of one feature should not impact judgments of the other. If, on the other hand, the representations of the two features are entangled, then biasing judgments of one feature should necessarily impact judgments of the other in the way predicted by the entanglement.

We therefore used electrical microstimulation in V1 to bias judgments of one feature (location) and measured the impact on judgments of the other feature (spatial frequency) under different

task certainty conditions. Thanks to the retinotopic organization of V1, we were able to induce a bias in the perceived location change by microstimulating V1 during the presentation of the second stimulus. The bias was predicted by the tuning of the stimulated site: when the receptive field recorded on the microstimulated channel was located to the right of the first stimulus, the monkey reported rightward location shifts more frequently, and vice versa (Supplementary Figure S4(A)). On trials when the monkey performed the spatial frequency task, microstimulation (which affects essentially only location judgments; Fig. S4(B)) should not have influenced choices if the representations were orthogonal. Consistent with the task interference hypothesis, microstimulation had a bigger impact on spatial frequency judgments on trials with low than high task certainty (Fig. 4C open symbols significantly deviate more from the diagonal line than filled symbols), suggesting that the representations of location and spatial frequency were less orthogonal under low certainty.

Discussion:

Humans and animals must flexibly switch between tasks to survive in an uncertain world. This flexibility has long been known to come at a cost. The cost of flexibility has been explained by a diverse set of ideas, which are mostly modeled and tested using behaviour (13-20). We aimed to unify and mechanistically explain the extensive body of behavioural literature that has explored this intriguing phenomenon.

We used a multidisciplinary approach to identify neural population reasons for the cost associated with switching between tasks. By grounding our investigation in the neural representations of tasks and their associated variables, we uncovered a task interference neuronal mechanism that unifies previous findings and has implications for our understanding of many cognitive processes.

Integrative approach reveals neuronal mechanism for complex behaviour

In our research, we presented a novel research paradigm combining artificial neural networks, monkey electrophysiology, and human psychophysics, to understand the neuronal mechanism for advanced cognition and complex behaviour. By constraining artificial neural networks using actual behavioural observations, we generate data-driven hypotheses about the neuronal dynamics that underlie the intricacies in real behaviour. Large-scale electrophysiology in behaving monkeys allows for these hypotheses to be empirically tested at a single-neuron level, providing a direct examination of the neural correlates of hypothesized cognitive processes. Human psychophysics complements this, extending the relevance of our findings beyond a single species and ensuring applicability to broader cognitive phenomena.

Combined, our approach holds potential to extract meaningful insights from complex data, a critical challenge in an era when datasets become increasingly high dimensional (39).

Task interference may impact neural processes throughout the brain

Remarkably, we observed neural signatures of task interference even in V1, a primary sensory area not typically associated with strong modulations by cognitive processes (21, 22). This suggests that the phenomenon of interfering neural representations is not confined to specific

brain regions but rather extends throughout the brain. Future research should investigate whether task interference poses an even stronger limit in higher-order visual areas and prefrontal cortex, which are more intimately linked with more advanced cognitive processing.

Dimensionality as a fundamental limit on cognition

Limited capacity is a repeatedly reported constraint on multiple cognitive processes including working memory (23), attention (24), decision-making (25), and learning (26). Our findings raise the intriguing possibility that the dimensionality of a neural population representation is a fundamental limit on cognition because it limits the behavioural repertoire that the population can enable. The dimensionality limits the number of quantities (e.g. sensory features, number of objects, task variables, timescales, etc.) that can be represented independently. In theory, the dimensionality of a neural population representation could approach the number of neurons in that population, but the true dimensionality is typically much lower (27, 28).

Observations in diverse experimental systems are broadly consistent with the possibility that the dimensionality of neural population representations limits cognitive capacity. For instance, associating multiple simple features (such as the colour and surface texture of a banana) to aid in perceiving a complex feature (such as the ripeness of the banana), might combine, or reduce the dimensionality of, the representations of those associated features (29). However, reversing such learned associations would in this scenario require the brain to increase the dimensionality again, which could account for the challenges in reversal learning in operant conditioning in mice (30), rule learning in children with bipolar disorder (31), and language learning in neurotypical people (32).

Similarly, limitations in working memory capacity (23, 33) might be thought of as a limit on the number of items that can be represented independently in a population of neurons. People can overcome those limits by 'chunking' the memory of different items into larger, familiar units (e.g., memorizing the digits of a credit card or phone number in groups of three or four (33, 34)). This strategy may effectively reduce the neural dimensions needed to represent and maintain the whole sequence. Interestingly, artificial neural networks exhibit human-like working memory capacity limitations only after being pre-trained on large-scale natural image tasks (35), which may indicate that the mechanisms adopted to process high-dimensional inputs compromise performance on reduced laboratory tasks, perhaps due to an increase in interference between stimuli. The possibility that limitations on diverse cognitive processes share a neuronal mechanism offers a path toward understanding and alleviating those limitations.

Translational implications of task interference

Deficits in cognitive flexibility are an especially stark consequence of the many brain disorders, including dementia, substance abuse, and developmental disorders, that are associated with diminished cognitive capacity (36, 37). If task interference indeed underlies many cognitive limitations, our results offer hope for improving and repairing cognitive capacity. For example, we recently demonstrated that methylphenidate (trade name Ritalin, a stimulant used to treat disorders of attention; (38)) effectively increases the dimensionality of neural population representations in the visual cortex and that this increase is correlated with improvement in a visual task (38). This is a proof of principle supporting the idea that repurposing existing drugs to

address other cognitive abilities that might share a neural mechanism with their primary target could be an effective way to improve cognition in health and disease. The suggestion in our V1 results that task interference affects brain-wide mechanisms means that systemic drugs and/or therapies that affect widespread neuromodulatory systems might be especially effective.

Together, our results hold promising implications for the future of basic science and translational efforts to mechanistically understand, improve, and repair cognition. Our task interference hypothesis provides a concrete set of testable predictions with implications for theoretical and experimental work spanning many species and systems.

Methods:

Behavioural task for non-human primates

We designed a two-interval, two-feature discrimination task with stochastic task switching that allows us to measure and manipulate task certainty and perceptual decision-making (Xue et al, 2022). Briefly, two rhesus monkeys (both male, 12 and 9 kg) fixated a central spot while two Gabor stimuli (~2° visual angle in diameter) were displayed in series (200 ms each), separated by a variable delay (300-500 ms). The second Gabor stimulus differed from the first in its spatial location (shifted left or right) and spatial frequency (higher or lower), with independently randomized change amounts in the two features. Small titrations of the change amounts in the two features were made before each experiment to keep the perceptual performance of the animals in both tasks at roughly 75%. After a 150 ms second delay, the fixation dot disappeared, and the animals looked at one of four peripheral targets to indicate both the inferred relevant feature and the direction of change in that feature (spatial frequency increase or decrease and spatial location left-shift or right-shift). The relevant feature changed stochastically on 2.5% of trials, and the monkeys were rewarded only if they correctly discriminated the relevant feature. The visual display contained no information about the behavioural relevance of features, so the animals needed to infer the relevant feature based on their stimulus, choice, and reward history. The median number of trials in each daily session was 1086. All animal procedures were approved by the Institutional Animal Care and Use Committees of the University of Pittsburgh and Carnegie Mellon University, where the animal experiments were conducted.

Behavioural task for the human participants

We adapted our task for online human psychophysics (built with Gorilla experiment building tools (40) and customized scripts). We recruited a total of 220 adult human participants through Prolific (prolific.co), who performed slightly different combinations of behavioural experiments.

In the main experiments, we used a version of the two-interval, two-feature discrimination task with stochastic uncued task switching that we adapted to better suit human participants and online experiments (Figure 1B). The timing was slightly different (500 ms initial fixation period, 300 ms stimulus presentations, and 500 ms delay period). The two features to be discriminated were orientation and spatial frequency. As in the monkey experiments, the changes in the two features were independently varied. Similar to the monkeys, the change amounts were set at each participant's perceptual threshold, at which the participants performed perceptual

discriminations in both tasks at roughly 75% correct, determined by a psychophysical staircase procedure before the main task session. Participants indicated their choice by clicking on one of four buttons representing each feature and its change direction (indicated by a combination of text and graphical icon). As in the monkey experiments, the human participants received feedback following each choice (a green tick indicated a correct report of the direction of change in the task-relevant feature, and a red cross indicated an incorrect perceptual judgment, task choice, or both). As in the monkey experiments, the relevant feature switched stochastically, but the frequency was different (with a probability of 0.1 on each trial for humans, 0.025 for monkeys). The overall length of a typical session varied between 100 to 150 trials and around 15-20 minutes.

Task-belief and feature perception confidence report

Some task sessions required participants to report their confidence in their perceptual judgment on a random 20% of trials. After reporting their decision on the main task but before receiving feedback, participants were presented with three confidence sliders. Participants were instructed to indicate their task-belief on the first slider, ranging from a strong belief that spatial frequency was relevant to a strong belief that orientation was relevant. An intermediate value indicated low task certainty. The second and third sliders were used to report the perceived direction of change for each feature, where the absolute value reflected confidence in the perceptual judgment.

Variable delay and feature recall question

In some sessions, we explored memory for believed-relevant and believed-irrelevant stimulus features by interrupting a randomly selected 20% of trials after either 500 ms (short delay) or 2000 ms (long delay) to ask participants to recall a feature of the first Gabor. Participants indicated the feature memory using a slider, facilitated by four reference Gabor patches that spanned the possible range. We quantified estimation error as the difference between the actual stimulus feature and the recalled feature.

RNN models and analyses

Training and architecture

We trained RNN models using custom code based on PsychRNN (Ehrlich et al., 2021) and TensorFlow (Abadi et al., 2016). The models were governed by the following dynamical equations:

$$\tau\, dx = (-x + W_{rec}\, r + b_{rec} + W_{in}\, u)\, dt + \sigma_{rec} \sqrt{2\tau}\, d\xi$$
$$r = f(x) = max(0, x)$$
$$z = W_{out}\, r + b_{out}$$

where x is the recurrent state, which when passed through the rectified linear unit (ReLU) activation function gives the firing rate r, u is the input vector, and z is the output vector. W_{in}, W_{rec}, and W_{out} are the input, recurrent, and output weight matrices, and b_{rec} and b_{out} are constant biases added to the recurrent and output units. dt is the discrete time step, τ is the neuronal time constant, and σ_{rec} scales recurrent noise, which is a Gaussian noise process ($d\xi$).

Our monkey choice network (Figure 2A) was trained to predict the choices of two macaques based on the same history of stimuli, choices, and rewards experienced by our monkey

participants. The RNN had two modules, which we term the 'task module' and the 'perception module.' The two modules were fully connected in the recurrent layer but received different inputs and mapped onto different outputs. The task module received the history of monkeys' choices, stimulus feature change amounts (signaling perceptual difficulty), and rewards for nine preceding trials. These inputs were constant throughout the trial, based on the assumption that task rule dynamics are slow compared to single-trial stimulus response dynamics. The task module was trained to output the task that the monkey chose on the current trial as a one-hot vector. The perception module received the spatial frequency (SF) and spatial location (SL) values of the first stimulus and, after a variable delay (300 to 500 ms, uniform distribution), of the second stimulus. Units in the perception module displayed mixed selectivity for the two features. The perception module was trained to output one of the four choices (SF increase, SF decrease, SL decrease, SL increase) as a one-hot vector.

We compare this monkey choice network to a correct choice network (Figure 2A), which was trained to output the correct choice (rather than the monkeys' choices) given the perceptual inputs on the current trial and its own trial history (feedback based on the model's choices rather than the monkeys'). Stimulus values were drawn from the same data used to train the monkey choice network. After the models were trained, we used the monkeys' trial history inputs for both models so that we could directly compare trials (all model figures besides 2B). The correct choice network trace in figure 2B was calculated directly from its training batches (200 trials per epoch), and sliding window averaged over 100 epochs. The monkey choice network trace was calculated post-hoc from weights saved during training (every 15 epochs), and sliding window averaged over 6 datapoints (90 epochs). We chose to analyze the monkey choice network at the point during training with the highest monkey choice accuracy (see Figure S1), which was at training epoch 1245 out of 1500. All RNNs were trained using the Adam optimizer, with L1 and L2 regularization of weights and L2 regularization of firing rates. Note that to keep the model architecture consistent with the monkey choice network, the correct choice network here has key differences in the inputs from the model we used in our previous study (3).

Reaction time calculation

To calculate model reaction times, we applied a collapsing decision boundary of the form $f(t) = a \cdot \left(1 - \frac{b \cdot t}{t+c}\right)$ to the choice output units of the RNNs, where t is the time after choice period onset. For figure 2D, the parameter values were a=1.5 ms, b=1.8, c=28 ms.

Distance covariance analysis for assessing orthogonality of feature axes

To quantify the (non-)orthogonality of feature axes in the models, we used distance covariance analysis (DCA), a dimensionality reduction method that identifies linear projections that capture interactions between variables by maximizing the correlational statistic distance covariance (41). After reducing the dimensionality of the activity of all units in the hidden layer to the first 10 principal components (capturing >95% of the variance), we calculated the distance covariance statistic and corresponding axis separately for each unique set of trial history inputs, feature (location and spatial frequency), and timepoint during the delay period, from 500 simulated trials with different feature values. The projection of the irrelevant feature axis onto the relevant

feature axis was calculated by taking the dot product of the unit vectors and scaling it by the smaller distance covariance (that of the believed-irrelevant feature) (Figures 2F and S3).

Electrophysiology for the monkey experiments

Different aspects of a subset of the electrophysiological data have been reported previously (Xue et al, 2022). The visual stimuli were displayed on a linearized CRT monitor (1,024 × 768 pixels, 120-Hz refresh rate) placed 57 cm from the animal. We monitored eye position using an infrared eye tracker (Eyelink 1000, SR Research) and used custom software (written in MATLAB using the Psychophysics Toolbox, (25) to present stimuli and monitor behaviour. We recorded eye position and pupil diameter (1,000 samples per s), neuronal responses (30,000 samples per s), and the signal from a photodiode to align neuronal responses to stimulus presentation times (30,000 samples per s) using hardware from Ripple. We recorded neuronal activity from chronically implanted Utah arrays (Blackrock Microsystems; 48 electrodes in V1) during daily experimental sessions for several months in each animal (90 sessions from monkey F and 68 sessions from monkey G). We set the threshold for each channel at three times the standard deviation and used threshold crossings as the multiunit activity on that unit. We positioned the stimuli to maximize the overlap between potential stimulus locations and the joint receptive fields of V1 units. The receptive fields were measured using separate mapping sessions, during which the monkeys fixated their gaze on a dot, while Gabor stimuli were subsequently flashed at random locations for 100ms each across a portion of the screen.

We included experimental sessions for analysis if the total number of completed trials was at least 480 (which was ten times the number of channels on the multielectrode array). During complete trials the monkeys successfully maintained fixation until they indicated their choice). We analyzed the activity of V1 units during stimulus display periods, shifted by 34 ms neuronal response latency (42). V1 units are included for analyses if the stimulus response is 25% larger than baseline activity, measured 100 ms before stimulus onset, and 2) larger than 5 sp/s. These procedures resulted in 89 sessions from Monkey F and 68 sessions from Monkey G; an average 46 V1 units, and a median 1086 completed trials per session. Other aspects of these data have been reported in a previous study (3).

We did V1 microstimulation experiments during 6 sessions in one monkey (monkey F, Figure 4). We were unable to conduct these experiments in the other monkey because the experiments were interrupted by the COVID shutdown of the lab, after which there were no usable channels on the array. During microstimulation sessions, we electrically stimulated one channel on the V1 array with 20, 30 µA, 200 Hz biphasic pulses between 50 and 150 ms after the onset of the second Gabor stimulus on a randomly selected 50% of trials.

Microstimulation perturbed the monkey's reported location shift consistently based on the relative locations of the V1 receptive field and the first Gabor location (Supplementary Figure S4(A), but did not effectively manipulate the monkeys reported frequency change based on the preferred frequency relative to the first Gabor frequency (Supplementary Figure S4(B). This discrepancy may be attributed to the fact that microstimulation likely activates neurons through direct axonal activation within a volume of tens of microns in diameter (43). While these activated neurons may exhibit different frequency selectivity than the recorded neurons on the same channel, they likely share similar receptive fields due to V1's retinotopic structure. We

assessed the level of interference as the degree to which the receptive field location relative to the first Gabor location influenced the monkey's frequency choices (which should be zero if the representations of frequency and location are orthogonal). We then compared the level of interference induced by microstimulation following rewarded trials and following unrewarded trials.

Statistical tests

All p-values reported in this study are from Wilcoxon signed rank tests unless otherwise specified.

Acknowledgments

We are grateful to John Maunsell, Douglas Ruff, Ramanujan Srinath, and Pouya Bashivan for comments on this manuscript. This work was supported by the Simons Foundation (Simons Collaboration on the Global Brain award 542961SPI to MRC), the National Eye Institute of the National Institutes of Health (awards R01EY022930, R01EY034723, and RF1NS121913 to MRC).

Author contributions

C.X. and M.R.C. designed research; S.K.M. and C.X. trained recurrent neural nets and analyzed the data; C.X. and L.E.K. performed monkey training and neuronal recording; C.X. analyzed the neuronal data and performed the microstimulation experiments; R.C. and C.X. performed the online human psychophysics experiments and data analyses; M.R.C. supervised the findings of this work; C.X., S.K.M and M.R.C. wrote the paper.

References:

1. Monsell S. Task switching. Trends Cogn Sci. 2003;7(3):134-40.
2. Kiesel A, Steinhauser M, Wendt M, Falkenstein M, Jost K, Philipp AM, et al. Control and interference in task switching--a review. Psychol Bull. 2010;136(5):849-74.
3. Xue C, Kramer LE, Cohen MR. Dynamic task-belief is an integral part of decision-making. Neuron. 2022.
4. Koch I, Poljac E, Müller H, Kiesel A. Cognitive structure, flexibility, and plasticity in human multitasking-An integrative review of dual-task and task-switching research. Psychol Bull. 2018;144(6):557-83.
5. Longman CS, Lavric A, Munteanu C, Monsell S. Attentional inertia and delayed orienting of spatial attention in task-switching. J Exp Psychol Hum Percept Perform. 2014;40(4):1580-602.
6. Mayr U, Kuhns D, Rieter M. Eye movements reveal dynamics of task control. J Exp Psychol Gen. 2013;142(2):489-509.
7. Longman CS, Lavric A, Monsell S. More attention to attention? An eye-tracking investigation of selection of perceptual attributes during a task switch. J Exp Psychol Learn Mem Cogn. 2013;39(4):1142-51.
8. Pashler H. Dual-task interference in simple tasks: data and theory. Psychol Bull. 1994;116(2):220-44.

9. Allport A, Wylie G. Task-switching: Positive and negative priming of task-set. 1999.
10. Yeung N, Nystrom LE, Aronson JA, Cohen JD. Between-task competition and cognitive control in task switching. J Neurosci. 2006;26(5):1429-38.
11. Waszak F, Hommel B, Allport A. Task-switching and long-term priming: role of episodic stimulus-task bindings in task-shift costs. Cogn Psychol. 2003;46(4):361-413.
12. Allport A, Styles EA, Hsieh S. Shifting Intentional Set: Exploring the Dynamic Control of Tasks. 1994.
13. Ardid S, Wang XJ. A tweaking principle for executive control: neuronal circuit mechanism for rule-based task switching and conflict resolution. J Neurosci. 2013;33(50):19504-17.
14. Driscoll L, Shenoy K, Sussillo D. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. bioRxiv. 2022:2022.08.15.503870.
15. Yang GR, Joglekar MR, Song HF, Newsome WT, Wang XJ. Task representations in neural networks trained to perform many cognitive tasks. Nat Neurosci. 2019;22(2):297-306.
16. Mante V, Sussillo D, Shenoy KV, Newsome WT. Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature. 2013;503(7474):78-84.
17. Flesch T, Juechems K, Dumbalska T, Saxe A, Summerfield C. Orthogonal representations for robust context-dependent task performance in brains and neural networks. Neuron. 2022;110(24):4212-9.
18. Riveland R, Pouget A. Generalization in Sensorimotor Networks Configured with Natural Language Instructions. bioRxiv. 2023:2022.02.22.481293.
19. Deco G, Rolls ET. Attention and working memory: a dynamical model of neuronal activity in the prefrontal cortex. Eur J Neurosci. 2003;18(8):2374-90.
20. Duan CA, Pagan M, Piet AT, Kopec CD, Akrami A, Riordan AJ, et al. Collicular circuits for flexible sensorimotor routing. Nat Neurosci. 2021;24(8):1110-20.
21. Luck SJ, Chelazzi L, Hillyard SA, Desimone R. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. J Neurophysiol. 1997;77(1):24-42.
22. McAdams CJ, Reid RC. Attention modulates the responses of simple cells in monkey primary visual cortex. J Neurosci. 2005;25(47):11023-33.
23. Cowan N. Working memory capacity: Classic edition: Psychology press; 2016.
24. Pylyshyn ZW, Storm RW. Tracking multiple independent targets: evidence for a parallel tracking mechanism. Spat Vis. 1988;3(3):179-97.
25. Kang YH, Löffler A, Jeurissen D, Zylberberg A, Wolpert DM, Shadlen MN. Multiple decisions about one object involve parallel sensory acquisition but time-multiplexed evidence incorporation. bioRxiv. 2020:2020.10.15.341008.
26. Van Merrienboer JJ, Sweller J. Cognitive load theory and complex learning: Recent developments and future directions. Educational psychology review. 2005;17:147-77.
27. Saxena S, Cunningham JP. Towards the neural population doctrine. Curr Opin Neurobiol. 2019;55:103-11.
28. Chung S, Abbott LF. Neural population geometry: An approach for understanding biological and artificial neural networks. Curr Opin Neurobiol. 2021;70:137-44.
29. Libby A, Buschman TJ. Rotational dynamics reduce interference between sensory and memory representations. Nat Neurosci. 2021;24(5):715-26.
30. Meister M. Learning, fast and slow. Curr Opin Neurobiol. 2022;75:102555.
31. Wegbreit E, Cushman GK, Weissman AB, Bojanek E, Kim KL, Leibenluft E, et al. Reversal-learning deficits in childhood-onset bipolar disorder across the transition from childhood to young adulthood. J Affect Disord. 2016;203:46-54.
32. Idemaru K, Holt LL. Word recognition reflects dimension-based statistical learning. J Exp Psychol Hum Percept Perform. 2011;37(6):1939-56.

33. MILLER GA. The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol Rev. 1956;63(2):81-97.
34. Thalmann M, Souza AS, Oberauer K. How does chunking help working memory? J Exp Psychol Learn Mem Cogn. 2019;45(1):37-55.
35. Xie Y, Duan Y, Cheng A, Jiang P, Cueva CJ, Yang GR. Natural constraints explain working memory capacity limitations in sensory-cognitive models. bioRxiv. 2023:2023.03.30.534982.
36. Ferreira CS, Maraver MJ, Hanslmayr S, Bajo T. Theta oscillations show impaired interference detection in older adults during selective memory retrieval. Sci Rep. 2019;9(1):9977.
37. Dajani DR, Uddin LQ. Demystifying cognitive flexibility: Implications for clinical and developmental neuroscience. Trends Neurosci. 2015;38(9):571-8.
38. Ni AM, Bowes BS, Ruff DA, Cohen MR. Methylphenidate as a causal test of translational and basic neural coding hypotheses. Proc Natl Acad Sci U S A. 2022;119(17):e2120529119.
39. Fetsch CR. The importance of task design and behavioural control for understanding the neural basis of cognitive functions. Curr Opin Neurobiol. 2016;37:16-22.
40. Anwyl-Irvine AL, Massonnié J, Flitton A, Kirkham N, Evershed JK. Gorilla in our midst: An online behavioural experiment builder. Behav Res Methods. 2020;52(1):388-407.
41. Cowley B, Semedo J, Zandvakili A, Smith M, Kohn A, Yu B. Distance Covariance Analysis. In: Aarti S, Jerry Z, editors. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics; Proceedings of Machine Learning Research: PMLR; 2017. p. 242--51.
42. Schmolesky MT, Wang Y, Hanes DP, Thompson KG, Leutgeb S, Schall JD, et al. Signal timing across the macaque visual system. J Neurophysiol. 1998;79(6):3272-8.
43. Histed MH, Bonin V, Reid RC. Direct activation of sparse, distributed populations of cortical neurons by electrical microstimulation. Neuron. 2009;63(4):508-22.

Supplementary Figures

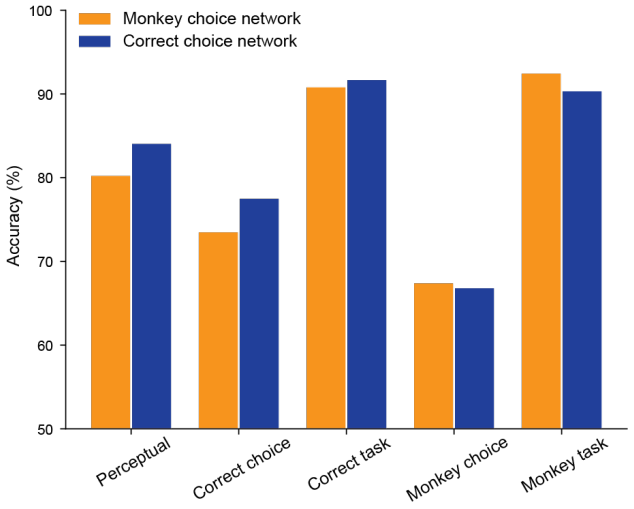

FIG S1: Accuracies of the monkey choice network (orange) and correct choice network (blue) after training. Perceptual: percentage of correct perceptual judgments, regardless of the chosen task (chance=50%); Correct choice: percentage of correct choices (chance=25%); Correct task: percentage of correct task choices, regardless of the perceptual judgment (chance=50%); Monkey choice: percentage of choices that match the monkeys' choices

(chance=25%). Monkey task: percentage of task choices that match the monkeys' chosen task, regardless of the perceptual judgment (chance=50%). All differences between models are statistically significant (two-sided z-tests, p<0.0005).

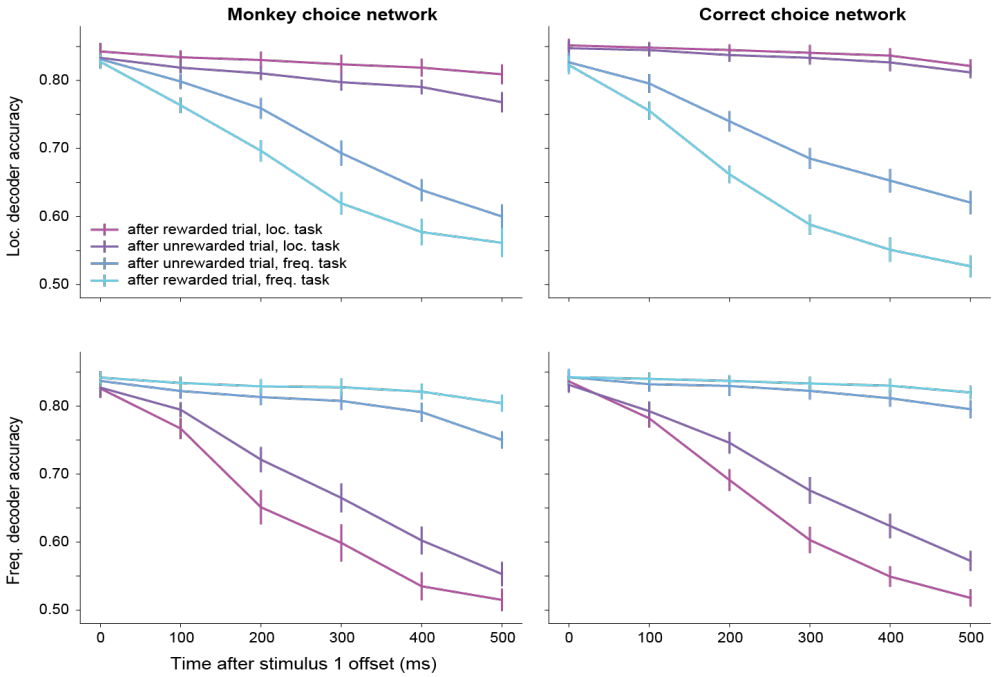

FIG S2: Feature (location, top row, and frequency, bottom row) decoder accuracies during a 500 ms interstimulus delay period for the monkey choice network (left column) and correct choice network (right column) in four trial conditions. The conditions are as follows: 1) after a rewarded trial, location task chosen (high location task certainty); 2) after an unrewarded trial, location task chosen (low location task certainty); 3) after an unrewarded trial, frequency task chosen (low frequency task certainty); 4) after a rewarded trial, frequency task chosen (high frequency task certainty). In both models, the more a feature is believed to be irrelevant, the more quickly information about that feature decays following stimulus offset. Each point represents the mean performance of cross-validated linear decoders (same as Figure 2E), with error bars showing +/- 1 standard deviation. A separate decoder was trained for each stimulus condition (identical first stimulus) and timepoint.

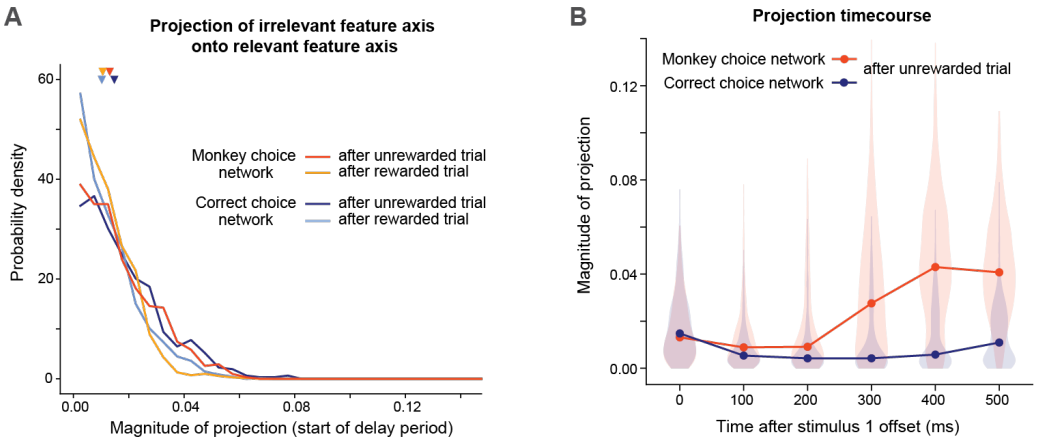

FIG S3: Feature axes become less orthogonal following stimulus offset in the monkey choice network but not the correct choice network. Orthogonality is quantified by calculating the projection of the irrelevant feature axis onto the relevant feature axis. (A) At the beginning of the interstimulus delay period (0 ms after first stimulus offset), the monkey choice network represents the two features (location and frequency) as orthogonally as the correct choice network both after unrewarded trials (p=0.16) and after rewarded trials (p=0.67, Wilcoxon rank sum tests). The median projections (triangle markers) are all below 0.02. This contrasts with Figure 2F, which shows the same histograms at the end of the delay period (400 ms after first stimulus offset). (B) A violin plot showing the time course of projections during a 500 ms interstimulus delay period. Points represent the medians of the projection distributions; shaded regions show kernel density estimates. Feature axes exhibit a partial 'collapse' in the monkey choice network, but not the correct choice network, such that the projection of the irrelevant feature axis onto the relevant feature axis increases significantly.

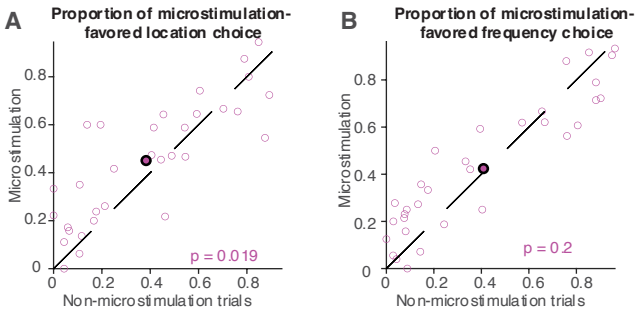

FIG S4: Electrical microstimulation induced location choice bias predicted by the receptive field (RF) of nearby recorded neuron (A) but did not significantly induce frequency choice bias predicted by the frequency selectivity of nearby recorded neuron (B). We selectively microstimulated V1 neurons during the second stimulus to observe the impact of microstimulation on behavioural choices. Each point in the scatterplot represents a stimulus condition. (A) In trials where the RF of the neuron found at the microstimulated site is to the left side of the first stimulus, then we define location left-shift choice as the "microstimulation favored location choice"; and vice versa for the location right-shift choice with RF to the right side of the first stimulus. The scatter plot displays the behavioural effects of microstimulation

on location choices by plotting the proportion of the microstimulation favored location choices with microstimulation (ordinate of the points) against that without microstimulation (abscissa of the points). The dots lie significantly above the diagonal line, showing that microstimulation biased the monkey's location choices predicted by the RF location of nearby recorded cells. (B) Similar figure convention to (A), except it shows the absence of a systematic behavioural frequency choice bias predicted by the frequency selectivity of nearby recorded cells