Title

Neural correlates of task switching in prefrontal cortex and primary auditory cortex in a novel stimulus selection task for rodents

Running Title

Neural correlates of auditory stimulus selection

Author List

Chris C Rodgers, 1 - xrodgers@gmail.com
Michael R DeWeese, 1, 2 - deweese@berkeley.edu
1 - Helen Wills Neuroscience Institute, University of California - Berkeley, 132 Barker Hall; Berkeley, California 94720
2 - Department of Physics, University of California - Berkeley, 132 Barker Hall; Berkeley, California 94720

Summary

Animals can selectively respond to a target sound in the presence of simultaneous distractors, similar to the way in which humans can respond to one person’s voice at a cocktail party. To investigate the underlying neural mechanisms, we recorded single-unit activity in primary auditory cortex (A1) and medial prefrontal cortex (mPFC) of rats selectively responding to a target sound from a mixture. We found that pre-stimulus activity in mPFC encoded the selection rule — the sound to which the rat would respond. Moreover, electrically disrupting activity in mPFC significantly impaired performance. Surprisingly, pre-stimulus and stimulus-evoked activity in A1 also encoded the selection rule, a cognitive variable typically considered the domain of prefrontal regions. However, stimulus tuning was not strongly affected. We suggest a model in which activation of a specific network of neurons underlies the selection of an imminent sound from a mixture, giving rise to robust and widespread rule encoding in both brain regions.
Highlights

1. Rats were trained on a novel auditory stimulus selection task
2. Pre-stimulus and stimulus-evoked activity in mPFC and A1 encoded selection rule
3. Stimulus selection did not strongly alter stimulus tuning in A1
4. Disruption of mPFC through electrical stimulation impaired task performance

Introduction

Humans can select and respond to one person’s voice even while many others are speaking at the same time. We do this effortlessly, yet no known algorithm can solve this “cocktail party problem” in realistic settings, perhaps because we do not fully understand the relevant computations performed in the brain (Cherry, 1953; Sayers and Cherry, 1957; Ding and Simon et al., 2012; McDermott, 2009). Other social animals such as birds and rodents demonstrate a similar ability (Bee and Micheyl, 2008); for instance, mother mice respond to distinct pup calls when several are calling at once (Geissler and Ehret, 2001). Humans use selective attention, the cognitive process of selecting and responding to a single target stimulus amongst simultaneous distractors (Desimone and Duncan, 1995), to solve the cocktail party problem (Ahveninen et al., 2011). Experiments in visual selective attention reveal that prefrontal cortex sends top-down “bias signals” to sensory cortex (Miller and Cohen, 2001; Moore et al., 2003) in order to select the target stimulus and subsequently enhance its neural representation, while suppressing the representation of distractors. Similar mechanisms may be at work in the auditory cortex: electrocorticographic (Mesgarani and Chang, 2013; Zion Golumbic et al., 2013) and magnetoencephalographic (Ding and Simon, 2012) recordings show that brain activity is dominated by the attended voice. Without recordings from single neurons it is difficult to ascertain what changes on the single-neuron level give rise to these effects.

Towards this goal, we have developed a new behavioral task for rats with three key properties. First, on each behavioral trial the subject hears a pair of simultaneous sounds, each drawn from a different category (e.g., white noise bursts vs. warbles). Second, the experimenter can indicate which sound the subject should select.
in order to receive a reward. Third, the subject then selects and responds to the correct sound from the pair. This means that the subject must be capable of selecting either of the two sounds upon demand, and in fact must be able to switch multiple times during a single behavioral session. This task requires cognitive flexibility because the same pair of stimuli demands a different behavioral response (“same stimulus; different response”) depending on which sound we require the subject to select. We are aware of no purely auditory single-unit studies in any animal satisfying these three conditions. The analogous ability in vision — to respond to a behaviorally relevant stimulus in the presence of competing distractors — has been referred to as stimulus selection (Knudsen, 2007; Reynolds and Chelazzi, 2004; Pestilli et al., 2011); following this, we refer to our task as auditory stimulus selection.

Similar visual and cross-modal tasks have been termed set shifting (Stoet and Snyder, 2004), task switching (Sasaki and Uka, 2009), and selective attention (Moran and Desimone 1985; Hocherman et al., 1976; Otazu et al., 2009). An alternative name for this type of task is stimulus feature selection, since two simultaneously presented sounds may be perceived as a single sound with two features. Other studies have investigated “response selection”: how decisions are translated into appropriate motor actions, following stimulus selection or even in the absence of an explicit stimulus (Young and Shapiro, 2011; Turken and Swick, 1999). We also note a similarity between our task and the Wisconsin Card Sorting Task for diagnosing disorders of executive function (Monchi et al., 2001). Our behavioral paradigm shares attributes with all of these, but for consistency we will refer to our task as stimulus selection below.

Although monkeys are the traditional model organism of choice for complex cognition (Gold and Shadlen, 2007), rodents are capable of sophisticated decision-making, in some ways very similar to humans (Raposo et al., 2012; Brunton et al., 2013; Zariwala et al., 2013). Rodents also show behavioral flexibility under the control of the prefrontal cortex (Karlsson et al., 2012; Kvitsiani et al., 2013; Young and Shapiro, 2011), even though this region is not necessary for simple sensory discriminations (Pai et al., 2011). The mPFC in particular appears to be critical for task switching (Birrell and Brown, 2000; Floresco et al., 2008; Durstewitz et al., 2010;
Ragazzino et al., 1999). For example, when rats learn to switch the navigational strategy they use to solve a maze, the mPFC encodes this switch and inactivating the area severely disrupts performance (Rich and Shapiro; 2009). Rodent mPFC thus appears to maintain a representation of the current task rule, analogous to the rule-encoding neurons observed in primate PFC (Wallis et al., 2001; Asaad et al., 2000; Johnston and Everling, 2007), although large parts of the monkey PFC appear to be functionally and anatomically unique to primates (Wise, 2008).

Frontal areas have been shown to play an important role in directing flexible auditory processing in A1 (Fritz et al., 2010). In that paradigm, ferrets were trained to detect tones at a specific target frequency, which resulted in rapid task-related plasticity (tuning changes) in A1 and increased functional connectivity with frontal areas (Fritz et al., 2010; Fritz et al., 2003). Moreover, the ferrets could rapidly switch between different auditory tasks and the character of the observed tuning changes matched the demands of each task (Fritz et al., 2005, Fritz et al., 2010). These experiments shed light on the mechanisms of attending to acoustic frequency and revealed A1 to be surprisingly dynamic for a primary sensory area, but the behaviors used were not stimulus selection according to our criteria. In our task, stimuli are always presented simultaneously rather than sequentially, and a stimulus used as a distractor on one trial can be the target on subsequent trials.

We are unaware of any single-unit studies of purely auditory stimulus selection in any animal. A model of this ability in rodents would be especially useful because of the relative ease and speed with which they can be trained on cognitively demanding tasks (Carandini and Churchland, 2013) and as a first step toward more complex behaviors, such as selective attention, which has traditionally been studied only in primates. Finally, many models of visual selection are not obviously applicable to the auditory modality — for instance the idea that visual attention co-opted the neural mechanisms for shifting gaze over evolutionary time. Establishing an auditory selective attention paradigm could shed more light on whether the known mechanisms of visual selection are universal or specific to one modality.
We recorded from individual neurons in both mPFC and the primary auditory cortex (A1) of rats performing our task. We found that the pre-stimulus, anticipatory activity of our recorded neurons in mPFC encoded which sound would be selected. Surprisingly, we also found this pre-stimulus effect in a sizable fraction of our recorded neurons in A1. Finally, stimulus-evoked activity in both brain regions was similarly modulated, although this did not appear to alter tuning properties in a way that would be obviously beneficial for responding to the selected sound.

Results

A novel behavioral task for rodents: auditory stimulus selection

We developed an auditory stimulus selection task for rats, in which the subject was trained to respond to either of two simultaneously presented sounds.

The rat initiated each trial (Figure 1A) by holding its nose in the center port of a three-port behavior box for at least 250 ms — the “hold period.” This triggered speakers on the left and right to play in stereo one of the following four equally likely stimulus pairs: LEFT+HIGH, RIGHT+HIGH, LEFT+LOW, or RIGHT+LOW (Figure 1B). Each stimulus pair was a simultaneous combination of a broadband noise burst from either the LEFT or RIGHT speaker, and either a HIGH- or LOW-pitched warble (frequency-modulated tone) from both speakers simultaneously. After the onset of stimulus presentation, the rat could then choose to “go left” (poke its nose in the left port), “go right” (poke its nose in the right port), or “nogo” (not poke either side). Correct pokes into the side ports were rewarded with water; incorrect pokes were penalized with a 2-6 s timeout (Methods).

On each trial, one of the sounds in the stimulus pair (the “target”) indicated the correct response but the other sound (the “distractor”) was uninformative. To indicate which sound the rat should select, the behavioral session alternated between “localization” blocks of trials (during which the noise burst was the target) and “pitch discrimination” blocks (during which the warble was the target). Each block consisted of 80 trials (Figure 1C), the first 20 trials of which were reserved to indicate the block change. During these 20 “cue trials,” the rat
heard only target sounds without any distractor. Behavioral controls (Figures S2B, S2C) showed that the rats responded to the target sound, not to the target/distractor combination.

We refer to this task as auditory stimulus selection, by which we mean that the rat selectively responds to one of two simultaneous sounds on any given trial. Importantly, a stimulus selection task requires the subject to be able to select either of the two sounds, depending on which one the experimenter designates as predicting reward. This designation could be accomplished with cues presented before the start of each trial, but in our task we use a block design with no explicit pre-trial cue, so the rat must use its recent reward history to determine which sound it should select.

Stimulus selection — selectively responding to a behaviorally relevant target in the presence of distractors — is one component of selective attention, a broader and more complex ability that also includes perceptual enhancement (Knudsen, 2007; Reynolds and Chelazzi, 2004; Pestilli et al., 2011). We feel that tasks requiring the detection of faint stimuli (Cohen and Maunsell, 2009) are the gold standard of selective attention research because they demonstrate perceptual enhancement (i.e., lower perceptual threshold) when the subject is directed to the task-relevant stimulus compared to trials when the subject attends to a distractor or is not directed to any stimulus. Although our task does not require a lowered perceptual threshold, it still represents an important step forward: we are aware of no other rodent paradigms for studying stimulus selection, nor any single-unit studies of purely auditory stimulus selection in any animal.

**Rats perform the task well above chance**

We ensured that the rats were in fact selecting the correct target sound by verifying that their behavioral response was driven by the target sound, significantly above chance and significantly more than it was driven by the distractor sound, and also that they were not using the same stimulus/response strategy in both blocks. Some strategies allow 50% performance without using any information from the target, such as always going to the choice port for the current block even in response to a nogo target, a strategy that we commonly observed in rats before they were fully trained. For this reason we verified that performance was significantly
and consistently greater than 50% in both blocks, and also that the animals were responding to the target sound and not the distractor (Methods), before and after implanting the recording electrodes. This typically required about 40 one-hour training sessions, for up to eight weeks. Our best rats’ typical performance during recording sessions was approximately 85% in both blocks (Figure 2). In general, the rats performed well above chance, rapidly and correctly changing which sound they selected after each block change. Rats performed both localization and pitch discrimination well, though they typically did better on go than on nogo trials (Figure S2A).

Because our task associates a different choice/reward port with each block, a different distribution of motor responses is required for localization (50% go left, 50% nogo) than for pitch discrimination (50% go right, 50% nogo). We note two consequences of this. First, this allows us to identify an interesting type of error trial on which the rat appeared to respond to the wrong sound. On such “interference” trials, the rat heard a “go” distractor (i.e., a sound to which the rat should respond with a go response in the other block) and incorrectly went to the choice port associated with that distractor, instead of doing what the target sound indicated. We analyze the neural correlates of this error below. Second, it is plausible that the rat’s motor plan differs between the blocks. There is a similarity in this sense between our task and some blocked visual spatial attention tasks, in which 80% of the trials require a saccade in the same direction (Cohen and Maunsell 2009). It can be difficult to tease apart response selection from stimulus selection (but see Erlich et al., 2011; Sato and Schall, 2003; Steinmetz and Moore 2012). We return to this issue later.

Anticipatory neuronal activity in mPFC encodes the selection rule

We next asked what differences in neuronal activity between blocks correlated with the selection of the target. We implanted tetrodes into the brain, targeting A1 and/or the prelimbic region of mPFC, and recorded single-unit action potentials (spikes) from multiple neurons during behavior. By analogy with the rule-encoding neurons in primate prefrontal cortex, we hypothesized that mPFC would encode the selection rule. That is, we expected that the firing rates of single mPFC neurons would differ significantly between localization and pitch discrimination trials. We first confined our analysis to the hold period, the interval before stimulus onset while
the rat is holding its nose in the center port and presumably preparing to select the target sound from the imminent stimulus pair.

We found that the hold period activity of a majority of mPFC neurons robustly encoded the selection rule on correct trials. An example unit (Figure 3A) fired significantly more \((p < 0.001, \text{Mann-Whitney U-test})\) in the hold period during localization trials (mean: 12.1 Hz) than it did during pitch discrimination trials (mean 7.2 Hz). A different but simultaneously recorded single unit in mPFC (Figure 3B) fired significantly more during pitch discrimination (mean 5.4 Hz) than during localization (mean 2.7 Hz). In both cases the effect persisted across the entire session of over 1300 trials, alternating with each block just as the behavior did. Across our recorded population of mPFC neurons, 63\% (76/121) of the neurons individually and significantly encoded the selection rule during the hold period (Figure 3C). Of these, 36 neurons preferred \(i.e.,\) fired more during localization and 40 preferred pitch discrimination; neither preference was significantly more common (binomial test, \(p > 0.05\)).

**Anticipatory neuronal activity in A1 also encodes the selection rule**

Surprisingly, we also found a similar effect in A1 (Figure 4). Although encoding of selection rule was our hypothesized result in mPFC, this was unexpected in A1, especially given the absence of auditory stimulation in the pre-stimulus period. Across our recorded population, 36\% (36/99) of A1 neurons encoded selection rule. As with mPFC, neither population was significantly larger (13 preferring localization, 23 pitch discrimination; binomial test, \(p > 0.05\)). Since A1 is known to encode many types of sounds in a sparse fashion (DeWeese et al., 2003; Hromádka et al., 2008; Carlson et al., 2012), we were not surprised to observe that only some of our recorded neurons in A1 significantly responded to our task stimuli (Supp. Info.). However, rule encoding was approximately equally widespread in both stimulus-responsive (14/49) and non-responsive (22/50) neurons. This finding is reminiscent of human imaging results suggesting that neurons in auditory cortex may carry top-down attention signals even in the absence of stimulus information (Ahveninen et al., 2011).

These effects were strong: among the significantly rule-encoding neurons, the median increase in firing rate during the preferred block was 74.7\% in mPFC and 99.7\% in A1. We controlled for the possibility that these
results in either brain region could be explained by firing rate drift over the course of the session or by spike sorting errors arising from small differences in spike waveform shape between blocks (Supp. Info.). We did not observe clustering or any other topographic organization of neurons preferring the same block, which implies that these effects could have been obscured had we relied on multi-unit recordings. In sum, these results demonstrate widespread and robust encoding of selection rule in the pre-stimulus activity of both mPFC and A1 neurons.

**Postural contributions to neuronal effects**

The mPFC regulates cognitive state, but it also plays a role in motor planning (Erlich et al., 2011). A classic study (Euston and McNaughton, 2006) showed that PFC neural activity, apparently related to working memory, might be well-explained solely in terms of behavior variability. Specifically, if neurons are selective for body position and body position differs between blocks, then encoding of rule may in fact be explained as encoding of body position. We analyzed video of our rats and found evidence of preparatory changes in head angle that differed between blocks (Supp. Info.). However we found that the block rule explained more of the neural variability than head angle did in the vast majority of rule-encoding neurons (Figure S3I-L, S4I-L). Moreover, as we discuss in the section after next, the long duration of the neural effects we observed in both brain regions further argued against the possibility that changes in posture were the underlying cause. Nonetheless, the role of motor planning remains an important consideration for interpreting recordings from prefrontal cortex and even, given our results in A1, sensory cortex.

**Error trial analysis**

In the previous sections we considered only correct trials. We next considered interference trials, during which the rat erroneously chose the port associated with the other block, suggesting that it was selecting the wrong sound from the mixture. If encoding of selection rule in the anticipatory activity is important for successful stimulus selection, then the encoding should be weaker or even reversed when the rat selected the wrong sound.
Indeed, in mPFC the encoding of selection rule was significantly weakened on interference trials, versus correct trials (Figure 3D). In A1, we observed a more extreme effect (Figure 4D): the rule encoding was actually reversed on interference trials — firing rates were greater during the non-preferred block on such trials. These observations are consistent with the idea that anticipatory activity predicts which sound the rat will respond to, even for trials on which the rat appears to respond to the distractor by going to the wrong choice port. Although the activity thus predicts a motor response to the block-irrelevant port, it does not differ between trials on which the rat ultimately goes to the block-relevant port (correctly or incorrectly) or chooses the nogo response (Figures S3C, S4C).

**Within-trial timescale of the encoding of the selection rule**

We next asked how long before the stimulus the encoding emerged, and for how long afterwards it persisted. For each rule-encoding neuron, we compared across blocks the smoothed firing rates in every 50 ms bin before and after the stimulus onset, up to plus or minus 3 seconds from the stimulus onset. We thereby determined the largest interval of time around the hold period during which the neural activity significantly encoded the selection rule. Across the dataset, the median inter-trial interval was 4.0 s (inter-quartile range: 2.7 s to 5.3 s) and so this time range (plus or minus 3 s) will overlap with the previous and/or next trial in many cases.

The temporal dynamics of the encoding varied widely across neurons in both regions (Figures 5A, B). For some neurons, rule encoding was strictly confined to the hold period: their firing rates were modulated only during the immediate pre-stimulus period. Other neurons showed significant encoding at all time bins tested: their firing rates were persistently elevated throughout the preferred block. We found neurons spanning this range of timescales in both brain areas. In A1, the median rule-encoding unit first developed a significant block preference 0.55 s pre-stimulus (IQR: 0.15 to 1.2 s); in PFC the median was 0.625 s pre-stimulus (IQR: 0.34 to 1.0 s). Thus, the majority of rule-encoding neurons developed this property well before the rat initiated a trial by center-poking (Figure S5B).
We asked whether the rule encoding reflected an increased firing rate in one block (versus spontaneous), a decreased firing rate in the other block (versus spontaneous), or something else (for instance, a low spontaneous rate, an elevated rate for one block, and an even higher rate for the other block, which might reflect an encoding of task difficulty). Individual neurons exhibited a diversity of effects and we observed single units showing each of these possibilities (Figure 5A, B; S5C, D, E). However we found that, across the population of rule-encoding neurons, the firing rate was significantly higher than spontaneous during the preferred block and significantly lower than spontaneous during the non-preferred block (Figures S3F, S4F). These data are more consistent with a “balanced” model, in which different subpopulations are active in each block, than an “unbalanced” model, such as one in which the modulation occurs only in one block as a result of that task being harder for the animal.

To examine the typical dynamics within each population and to determine which brain area first encodes the selection rule, we averaged the normalized activity (mean: 0, variance: 1) of all rule-encoding neurons in both brain regions during their preferred block. On average, the population activity ramped up gradually before stimulus onset, over a timescale of several seconds, and then fell relatively quickly afterward (Figure 5C). The activity in mPFC was first significantly elevated 2.7 seconds before stimulus onset, while population activity in A1 became elevated 0.88 seconds before stimulus onset. That the effect occurs first in mPFC is consistent with its hypothesized role as the origin of top-down bias signals to sensory cortex (Miller and Cohen, 2001); however, we emphasize that the wide range of timescales within both regions, and the fact that only a small fraction of our dataset consists of simultaneous recordings from A1 and PFC, complicates a direct comparison between brain regions.

**Encoding of behavioral choice**

We found a prominent difference between the firing rates on go and nogo trials, beginning near the time the rat left the center port and continuing for several seconds (Figure 5D). During a typical rule-encoding unit’s preferred block, its firing rate remained elevated on nogo trials for several seconds, during which time the rat
was generally beginning the next trial. In contrast, on go trials, the typical unit’s firing rate rapidly fell, and in fact remained below its long-term mean firing rate for several seconds, during which time the rat was typically moving to the reward port and consuming reward.

One interpretation of this result is that rule-encoding is particularly important for producing the nogo response. In this view, once the animal perceives the go stimulus, the rule-encoding disappears and the go response is produced. Another interpretation is that the rule-encoding is persistent on nogo trials because the animal is already preparing to begin the next trial less than a second later, whereas on go trials the animal no longer needs to encode the rule because the rat has already made its decision and is already moving to the reward port to consume water. This is consistent with previous reports of enhanced encoding of nogo stimuli (Fritz et al., 2003; David et al., 2012).

**Changes in baseline activity correlate with similar changes in evoked activity**

Given that the pre-stimulus activity encoded the selection rule, we next assessed whether the stimulus-driven activity in A1 differed between blocks. We first defined the evoked response window of each neuron as the period of time after stimulus onset during which the firing rate was significantly elevated above the pre-stimulus rate (Supp. Info.). The evoked response on each trial was then defined as the number of spikes emitted during this window. We analyzed the mPFC neurons in the same way and found a population of neurons showing auditory responses to our task stimuli that were low-latency and tightly locked to stimulus onset, similar to A1 (Figure 6A, B). Such neurons were rarer in PFC than in A1, though not significantly so (PFC: 31/90, A1: 49/99; p > 0.05, Fisher’s exact test). Evoked responses were significantly weaker in PFC than in A1 (Figure S6A-E).

Based on our finding that the increased firing rate during the preferred block often persisted for a period of time after stimulus onset, we expected that the evoked firing rate would also be higher during the preferred block. In both regions, this is indeed the case: an increase in pre-stimulus firing rate during one block correlates with an approximately equal increase in evoked firing rate during the same block (Figure 6C, D; exemplar: Figure 4B).
21% (9/43) of A1 neurons and 24% (4/17) of PFC neurons showed a significant elevation of evoked response during their preferred block.

We next used an ideal decoder analysis (Methods, Figure 6E) to ask whether the recorded neurons encoded the identity of the noise burst or warble with greater fidelity in either block, either due to changes in stimulus tuning, the baseline elevation described above, or some other effect. We can decode the identity of both the noise burst and the warble from the evoked responses in A1 (n=57 neurons in 22 simultaneously recorded ensembles) or PFC (n=25 neurons, 13 ensembles). The A1 neurons provide significantly better responses from which to decode sound identity, probably due to their stronger responses and tighter stimulus selectivity compared with the PFC neurons. However, for both brain regions and both types of sounds, we cannot decode the sound any more accurately from the localization trials than from the pitch discrimination trials. Moreover, we did not observe any correlation between each neuron’s change in anticipatory activity and its tuning for the stimuli (Figures S6F, G), or any indication that some stimulus pairs (e.g., the pairs requiring opposite responses in each block) elicited a greater response or a stronger modulation of response (Figures S6H, I).

However, as discussed in the previous section, we did observe encoding of the rat’s behavioral choice at a later timepoint, around the time the rat exited the center poke. This typically occurred 90 ms after stimulus onset, whereas the auditory-evoked onset responses discussed earlier in this section typically occurred 10-30 ms after stimulus onset (Figure S6A-E) and did not encode behavioral responses.

We conclude that, although neurons showing an increased pre-stimulus firing rate in one block generally showed an equivalent increase in the evoked rate during the same block, these changes in evoked rate do not obviously improve the detectability of the target sound. However, we note that our ensembles of neurons provided useful information about the identity of both sounds, and the brain has access to a pool of neurons orders of magnitude larger than our recorded population. It may be that the problem faced by cortex in this task is not to maximize the information available about the stimuli from individual neurons, but rather a wiring
problem of how to flexibly re-route the relevant stimulus information to the relevant motor neurons at every block change.

**Disruption of mPFC significantly impairs task performance**

mPFC has been shown to be required for many task switching paradigms, which prompted us to ask whether it is required for our task. To answer this question, we developed an electrical disruption technique, inspired by transcranial magnetic stimulation (TMS) in humans (Dayan et al., 2013). We first implanted the mPFC of three trained animals (rats Z1, Z2, and Z3) with extracellular stimulating electrodes. On 20% of trials (“disruption” trials), we injected a 10Hz train of current pulses (see Methods and Supp Info) during both the hold period and the duration of the auditory stimulus. Such electrical stimulation drives an extremely rapid activation of nearby neurons, followed by a slower suppression of firing rates (Butovas et al., 2003; Logothetis et al., 2010) for a few hundred milliseconds. Thus, the primary effect of this approach is neither to silence nor activate the brain region, but rather to disrupt the normal firing rates and patterns. Moreover, because we did not observe any spatial clustering of neurons preferring one task or the other, it is unlikely that such microstimulation would preferentially activate neurons of either preference, even if our stimulation protocol were purely excitatory.

Typically, pharmacological agents such as muscimol are used to test whether neural activity in the affected region is necessary for a certain behavior. Electrical disruption allows much finer control over the strength and timing with which we can perturb the circuit. We were able to stimulate during a desired subset of trials, for a certain time range (*i.e.*, throughout the center-poke hold and auditory stimulus presentation), while sparing activity the rest of the time. This provided us with a statistically powerful within-session control.

Across all three animals, electrical disruption tended to impair performance (Figure 7) in both localization (mean impairment 5.4% in Z1, 12.4% in Z2, and 27.5% in Z3) and pitch discrimination (19.1% in Z1, 18.7% in Z2, 13.0% in Z3). This impairment was significant across sessions (*p < 0.05*, binomial test) for pitch discrimination in 3/3 rats (Z1, Z2, and Z3) and for localization in 2/3 rats (Z2 and Z3). Electrical disruption largely, though not exclusively, affected performance on nogo trials. All rats were impaired on pitch
discrimination nogo trials in almost all sessions (Z1: 6/6, Z2: 8/8, Z3: 7/8 sessions). Some rats also exhibited additional impairments: Z3 was impaired on localization nogo (8/8 sessions) and Z2 was impaired on localization go trials (Z2, 8/8 sessions). These effects were generally quite strong within individual sessions (Figure 7B) even though they varied between rats. Taken together, these data suggest that, in the absence of normal mPFC activity, each rat resorts to its default strategy (typically “always go”) in one or both blocks. Normal activity in mPFC is therefore important for good performance in our paradigm, but the strong impairment on nogo trials in particular made it difficult to ascertain whether stimulus selection in particular was impaired, as opposed to impulse control, or some other aspect of the task (Figure S7).

**A simulated network model demonstrates how modulation of anticipatory activity could solve the stimulus selection problem**

Our data suggest a simple model of how the brain might perform stimulus selection, which we have elaborated into a quantitative simulation as a proof of principle. The model: 1) requires only random stimulus tuning in A1; 2) does not require tuning changes or synaptic reweighting after the initial training phase; 3) uses only excitatory connections, consistent with the observation that most long-range projections in the brain are excitatory.

The model (Figure 8) consists of a population of N neurons in A1, randomly tuned for each of our four stimulus pairs. The activation of each A1 neuron was subject to additive Gaussian noise. The “sensory SNR”, defined as the ratio of the strength of this noise to the strength of the stimulus tuning, is a free parameter. Half of the neurons are arbitrarily assigned to each of the two tasks. Each subpopulation projects to two command neurons encoding the two possible behavioral responses during that block (e.g., go left and nogo). Each projection is trained to activate the correct command neuron using a least-squares fit constrained to use only positive (excitatory) weights. The actual behavioral choice is determined by which command neuron is the most active (“winner-take-all”). Analytical equations for this procedure are given in the Supplemental Information.
After the training phase, the synaptic weights are fixed and a new set of test stimuli are presented. To produce the block-appropriate response, a “task signal” is added to the activations of the neurons in the appropriate A1 subpopulation for the current block, as indicated by either the red or blue neurons in Figure 8. Because all feed-forward weights are positive, adding this task signal translates into an excitatory boost to the premotor neurons appropriate for that block. Thus, even without any synaptic reweighting, the model will tend to choose the response appropriate for the current block and stimulus. With 320 neurons, the network performs above 80% correct even with a signal-to-noise ratio (SNR) as low as 0.0625 (i.e., very weak sensory responses in each neuron relative to its internal noise). Increasing the network size can lower this SNR limit even further (Figure S8).

This demonstrates that anticipatory modulation can be part of a scheme that is capable of solving the task switching problem, even with weak sensory responses and “random but fixed” tuning in A1, in which each subpopulation consists of randomly tuned neurons that do not change their tuning between blocks.

**Discussion**

**Auditory stimulus selection: task switching between conflicting auditory discriminations**

When human listeners hear two simultaneous voices, they can attend to and selectively respond to either one. This is a complex ability, and our task models part of it — selecting and responding to one of two simultaneous sounds. Our subjects can voluntarily switch which sound they select, and do so at each block change within a single recording session. The rats learned the task with less than eight weeks of training and performed many trials per session (median: 698; inter-quartile range: 507 to 912). To our knowledge this is the first published example of rodents performing such a stimulus selection task in any sensory modality.

Previous studies have identified critical roles for mPFC in behavioral flexibility in several contexts. For example, elegant work (Rich and Shapiro, 2009) established not only that the mPFC encodes the switches in
navigational strategies ("go east" vs "turn right") that rats use to solve a maze, but also that inactivating this region severely and selectively impairs their ability to perform the switch. Other studies of task switching in rodents required them to switch between a sensory discrimination and a (potentially habitual) fixed response ("follow the light" vs "always go left"; Floresco et al., 2008; Durstewitz et al., 2010). Many researchers are interested in extending these results to task switching between sensory discriminations, but it is often challenging to induce the switch when it requires ignoring a previously trained stimulus. Even in cross-modal switching, where the targets and distractors come from entirely different modalities, strong cueing mechanisms (violating our "same stimulus; different response" condition) have been used to induce the switch: introducing novel stimuli (Birrell and Brown, 2000), deleting distractors (Otazu et al., 2009), or changing the behavioral arena completely (Haddon and Killcross, 2007). Finally, most previous studies required rats to shift no more than once per session, sometimes just once per lifetime, while our study requires multiple switches per session. We believe our task advances the study of task switching in rodents to be much closer to the standard set by human and non-human primate studies.

Despite its clinical and computational relevance (Ding and Simon, 2012), the auditory cocktail party problem remains less studied than comparable visual tasks. Even in primates we are not aware of any single-unit studies of purely auditory stimulus selection. A multi-unit study (Lakatos et al., 2013) required monkeys to sustain attention to streams of pure tones; however, the researchers found that the monkeys were unable to ignore the distractor stream if it was within 1.5 octaves of the target stream. Human voices, even those with very different pitch, are typically much closer than this and actually overlap extensively in acoustic frequency (McDermott, 2009). For this reason, we believe animal models of this ability should use stimuli that, like ours, overlap at least partially in frequency and require solutions not based purely on frequency separation. In sum, we believe our task represents an important first step toward understanding the cocktail party problem in rats, paving the way toward future studies with the modern tools available in rodent models (e.g., the use of viral vectors expressing light-gated ion channels in specific brain regions or genetically-identified cell types).

**Anticipatory activity in both mPFC and A1 encodes the selection rule**
We found that rodent mPFC robustly encode the subject’s selection rule, analogous to the rule-encoding role of primate prefrontal cortex (Asaad et al., 2000; Wallis et al., 2001; Johnston and Everling, 2007). Rule encoding develops in our recorded mPFC population over 2.5 seconds before the stimulus onset, as the rat is planning to initiate a trial or even finishing the previous trial. The widespread nature of the encoding and the long timescales over which it persists are perhaps surprising because only one bit of information needs to be encoded — pitch discrimination or localization — and this information is only necessary while making a decision on each trial. One possibility is that this persistent activity represents a memory trace of the selection rule (Funahashi et al., 1989), meaning that it densely and persistently encodes cognitive variables like selection rule. In fact, the cortex may shift to a completely different network state (Karlsson et al., 2012) depending on which stimulus the rat plans to select.

We also observed anticipatory encoding of the selection rule in primary auditory cortex (A1), a surprising result since encoding of selection rule in the absence of sensory stimulation has traditionally been considered the domain of prefrontal areas. However, attention is known to modulate the pre-stimulus activity of single neurons in monkey V2 and V4, although not in V1 (Luck et al., 1997; Reynolds et al., 2000). At a larger spatial scale, visual attention can produce a similar increase in pre-stimulus baseline in V1, as assessed both with fMRI in humans (Pestilli et al., 2011) and with voltage sensitive dye in monkeys (Chen and Seidemann, 2011). Higher visual cortex also shows pre-stimulus modulation by attention in humans (Pestilli et al., 2011; Kastner et al., 1999; Thut et al. 2006). More generally, single neuron activity in primary sensory cortex can anticipate reward (Shuler and Bear, 2006) or a motor response (Niwa et al., 2012), and anticipation of a visual stimulus can trigger a hemodynamic response in V1, though without a corresponding change in neural activity (Sirotin and Das, 2009). In this light, perhaps it is not surprising that primary sensory cortex could also encode the selection rule for an imminent stimulus. In this way both the information about the stimulus and the information about how that stimulus should be interpreted are encoded in the same neurons, providing a possible locus for the behavioral decision to be made.
We observed a surprising amount of similarity between A1 and mPFC, both of which showed robust encoding of the selection rule and of behavioral choice (Figure 5D). In monkeys, attention effects become more prominent higher in the visual hierarchy (Luck et al., 1997). In contrast, our results show that rat A1 already robustly encodes a non-sensory variable, very similar to mPFC. This could be a difference between rats and monkeys, or between auditory and visual cortex, or both. Disambiguating these possibilities will be an important direction for future work.

**Comparison with studies of selective attention and task-relevant plasticity**

This pre-stimulus change in baseline contributed in an additive way to the strength of the sensory-evoked responses in both A1 and PFC. However, we found limited evidence for any additional modulation of sensory-evoked responses in A1. For example, the neurons did not appear to encode the target stimulus with any greater fidelity than the distractor stimulus. This is consistent with some, but not all, previous studies of auditory task switching. Although neuronal activity in A1 is robustly modulated in the aroused/engaged behavioral state versus the passive/idle state (Otazu et al., 2009; Lee and Middlebrooks, 2011), the neuronal effects of shifting between different engaged behaviors tend to be weaker or even non-existent. For instance, switching between an auditory task and an olfactory or visual task does not change evoked spiking auditory responses in A1 (Otazu et al., 2009; Lakatos et al., 2009), and switching between temporal and spatial auditory discriminations does not significantly change spatial tuning in A1 (Lee and Middlebrooks, 2011). Nonetheless, the fact that these studies (and ours) found no evidence of tuning changes in A1 does not mean that they do not exist under some circumstances.

In fact, a series of pioneering experiments demonstrated task-relevant plasticity in A1 of ferrets trained to detect a target frequency (Fritz et al., 2003; Fritz et al., 2010). One important methodological difference is that their study, unlike ours, made use of a large battery of probe stimuli and was therefore optimized to detect receptive field changes, including those affecting only task-irrelevant stimuli. Intriguingly, this plasticity was nuanced: it could induce facilitation or, alternatively, significant suppression at the task-relevant frequency.
Facilitation was more common than suppression, but the use of a different reinforcement paradigm reversed this (David et al., 2012). More studies of complex auditory behaviors will be necessary to better understand the factors that determine whether a given behavioral paradigm produces task-related modulation of evoked spiking responses in auditory cortex.

The lack of evidence for tuning modulation in our data is a surprising result, given that visual selective attention enhances target representations and suppresses distractors in V4 and other visual areas (Cohen and Maunsell, 2011; David et al., 2008; Mitchell et al., 2007; Reynolds and Heeger, 2009). However, selective attention consists of two component processes with separate behavioral measures: stimulus selection and perceptual enhancement (Knudsen, 2007; Reynolds and Chelazzi, 2004; Pestilli et al., 2011). Target-enhancing modulation of evoked responses is believed to mediate perceptual enhancement (although see Zénon and Krauzlis, 2012), as assessed behaviorally by a lower threshold or steeper psychophysical curves (Cohen and Maunsell, 2009; Moore et al., 2003). This suggests that only tasks that require perceptual enhancement might produce such effects.

In contrast, stimulus selection is often investigated with easily detectable stimuli far above threshold (Hocherman et al., 1976; Stoet and Snyder, 2004) and such studies, like ours, often report no modulation of evoked responses in sensory cortex (Sasaki and Uka, 2009; Mante et al., 2013). It may be that stimulus selection is the dominant computational challenge in such tasks and perceptual enhancement is therefore less important. Similarly, the cocktail party problem is often difficult because all voices are of competing intensity, not because the target voice is barely audible. Additive, pre-stimulus baseline increases have been observed in V1 during attention-demanding tasks and may lead to efficient stimulus selection (Chen and Seidemann 2012; Pestilli et al., 2011); our data support a similar hypothesis in A1. In summary, while the mechanisms by which selective attention mediates perceptual enhancement remain an important area of inquiry, enhancement of sensory evoked responses in A1 may not be necessary for our task or other similar stimulus selection paradigms.
The potential roles of motor planning and posture

We considered the potential roles of both posture — the angle of the rat’s head relative to the behavior box in particular — and motor planning in driving the observed effects. Because each block is associated with a different choice port, it is plausible that the rat adopts a different default motor plan for the two blocks — go left for one task and go right for the other. Moreover, we observed a difference in head angle between blocks, presumably a behavioral strategy that the rat used to prepare for the differing motor actions required.

We found that some rule-encoding neurons, especially in the mPFC, also encoded head angle to some extent. This is consistent with previous mPFC data (Euston and McNaughton, 2006) and the idea that single prefrontal neurons simultaneously encode disparate sensorimotor and cognitive signals (Rigotti et al., 2013). However, we found that the firing rate of most neurons was better explained by block than by head angle (Figures S3I-L, S4I-L) using a multiple-regression/ANOVA approach that our simulations showed to be well-suited for this problem (Figure S3Q-T). In addition, the rule encoding was largely maintained on a subset of “posture-equalized” trials, selected so that the mean head angle was the same in each block (Figure S3M-O, S4M-O). These results favor a hypothesis in which cognitive context (i.e., task rule) drives both the observed neuronal activity and the adaptive posture, over a hypothesis in which posture directly drives the observed neuronal activity.

To what extent do the observed effects reflect a motor plan, rather than a postural difference? It is difficult to separate motor planning from rule encoding because the task itself requires different sensorimotor mappings in each block. Moreover, a motor plan may be covertly present even in the absence of any measurable behavioral parameter, such as head angle. However, the timecourse of the neural effects we observed was quite protracted in many of our recorded neurons, in some cases even persistent throughout the block (Figure 5). During this period, the rat was engaged in various motor actions, such as moving to or from the center port and harvesting rewards in some cases (example neurons: Figure S3P, S4P). It seems unlikely that neurons would continue to represent the specific action of moving from the center port to the choice port on such a long
timescale. Note also that, for our paradigm, any default motor plan is subject to cancellation on nogo trials (sometimes called “countermanding”, cf. Schall et al., 2000; Eagle and Robbins 2003; Eagle and Robbins 2008), since the animal does not know during the anticipatory period whether it will be signaled to perform a go response or not. Finally, we did not observe any correlation between the anticipatory firing rate and the reaction or movement time (Figures S3G, H; S4G, H).

To summarize, our task requires remapping sensory stimuli to motor responses, and it is reasonable to expect rule encoding to incorporate both the sensory and motor planning aspects of this remapping.

**Stimulus selection via activation of latent circuits for each target**

Based on our results, we propose a model for stimulus selection based on task-specific activation of latent circuits, rather than task-specific adaptation of a single circuit. We found subpopulations of neurons in both A1 and mPFC — one activated during the localization block, the other during the pitch discrimination block. The signature of this activation is increased baseline activity. However, our recorded neurons do not change their tuning for specific stimuli. We hypothesize that the difference between the circuits is their downstream connectivity: each circuit may project to separate circuits in a downstream effector region, perhaps the striatum since the corticostriatal projection plays an important role in auditory decisions (Znamenskiy and Zador, 2013).

In this model, only one circuit is activated at a time, via feedforward excitation perhaps originating in mPFC, and only this circuit has sufficient baseline activity to drive behavior.

Our model makes several testable predictions. First, there should exist “premotor” neurons (possibly in the striatum) receiving input from A1 that also show a block-dependent anticipatory modulation. Second, neurons in A1 and in striatum showing the same block preference should be more strongly connected than those showing the opposite block preference. Finally, specific activation of one of the subpopulations in mPFC, A1, or striatum should specifically bias behavior toward the block preferred by that subpopulation. However, such a manipulation would require a means of stimulating only those neurons that can be functionally identified by their anticipatory firing rate, perhaps by expressing light-gated ion channels in the appropriate populations.
Possibly activity-dependent promoters such as cFos or other immediate early genes could be used to this effect.

In some ways, this model is more parsimonious than the traditional tuning change model of auditory attention, which requires that prefrontal (or other) brain regions be able to modulate the tuning of many A1 neurons as quickly as the subject shifts the focus of attention. Although attention does produce tuning changes (David et al., 2008; Fritz et al., 2003) over minutes (which is the fastest that they can be estimated from those data), it is unclear how known synaptic mechanisms could mediate task-specific tuning changes on a sub-second timescale. By contrast, our model requires only circuits with essentially fixed stimulus tuning, and the selection mechanism occurs by activating one of these circuits, rather than by changing the tuning of any of the neurons. This reflects the challenge of the task, which does not require amplifying the neural representation of a faint stimulus but rather a discrete change in sensorimotor mapping.

Taken together, our results are consistent with a distributed processing model in which contextual information from PFC modulates activity in A1 in order to increase the fidelity with which the appropriate motor action can be read out. This idea was proposed in the context of tuning changes (Fritz et al., 2010; David et al., 2012; Blake et al., 2002), but we demonstrate that it could also operate by activating a separate circuit without retuning neurons to task-relevant stimuli. Alternative models based on visual selection (Gilbert and Shallice, 2001; Mante et al., 2013) propose that stimulus selection occurs in frontal areas, not sensory cortex. Our data are similar to theirs in the sense that we do not observe tuning changes in sensory cortex (Mante et al., 2013) but different in the sense that we do not observe strong representations of the stimuli in PFC, similar to a recent observation in primate PFC (Lara and Wallis, 2012). We do observe encodings of the motor choice in both areas (Figure 5D). Whether these differences reflect a distinction between auditory and visual processing, or can perhaps be unified, remains a question for future work.
Our results establish the rat as a model organism for auditory stimulus selection, paving the way for future investigations of the cocktail party problem with emerging optical and genetic tools amenable to rodents. We have presented what we believe to be the first single-unit results in any animal performing an auditory stimulus selection task and we have found widespread and robust rule encoding in mPFC and A1, though we observed little change in the stimulus tuning of evoked responses. We propose a simple model to explain these results: task-specific activation of latent circuits, rather than task-specific adaptation of a single circuit.

Methods

Behavior training

We used male Long-Evans rats and began training them when their body mass reached 150g-200g, approximately 45-60 days old. Rats were given restricted access to water in the day before the training session so that they would be motivated to obtain water rewards. After each session they were given ad lib access to water for one hour. We monitored body weight and other markers to ensure they remained healthy.

We used a typical "shaping" procedure to train the rats. First they learned the localization task and pitch discrimination tasks separately and without a distractor. Next they learned to alternate between the tasks. Finally they learned to respond to the mixed stimulus containing target and distractor based on the block. Human intervention was required to determine when the rats were ready to progress to the next stage of training (generally, at least 80% hit rate). Human intervention was also required to discourage certain unwanted response strategies using the following tools: 1) increasing error timeout; 2) temporarily enforcing "all GO" or "all NOGO" trials (and dropping such trials from analysis); 3) giving water rewards out of the left or right port even in the absence of good performance in order to maintain motivation or encourage a task switch. Once the rats were sufficiently well-trained that little or no human intervention was required, they were implanted with the drive. Some rats required “retraining” after implantation using the techniques listed above; any trials thus affected were discarded from analysis. The entire training process takes about 10 weeks.

Trial timings
In three rats (Rats 1-3) the hold period was drawn from a uniform distribution on 0-100 ms; after pilot results indicated pre-stimulus effects, the hold period duration was increased to 250-350 ms in the other three rats. All trials with a hold period <50 ms were discarded for the analyses in Figure 3 and Figure 4. Hold period response was counted in the minimum window that applied to all trials: 50 ms for the first 3 rats and 250 ms for the rest.

The duration of the choice period differed between sessions, but was fixed within a session (or if it was changed slightly within a session, then the trials before the change were discarded from analysis). Correct entries into the choice port on go trials were rewarded with water from the same port. Incorrect entries into the choice port on nogo trials results in a 2-6 s timeout. Correct nogo responses were not explicitly rewarded with water, although the rat avoided a timeout with this response. Poking neither port on a go trial was not explicitly punished with a timeout, other than a lost opportunity for reward.

**Chance performance on the task**

In order for the rat to perform significantly above chance within a session, its behavior had to satisfy three criteria: 1) the rat performed significantly above 50% in each block, meaning that it must be using some information from the target sound (which is the only possible source of information on the correct response) to decide whether to go or nogo; 2) the rat is significantly more likely to perform the action indicated by the target than the action indicated by the distractor; 3) the rat is not using a “fixed strategy”, that is, the same mapping from stimulus pair to behavioral response in each block. (Because the target and distractor swap roles in each block, satisfying the second criterion is sufficient to satisfy the third.)

The first criterion rules out strategies like “always go left during localization”, which was a common strategy while first learning the task. We used a binomial test to compare the proportion of hits to 0.5 in both blocks and discarded any sessions that were not significantly (p > 0.05) above 50% in either block. The second criterion rules out certain hypothetical strategies such as always getting the congruent-nogo stimulus (RIGHT+HIGH) correct, and otherwise guessing randomly between the correct choices for that stimulus pair in each block.
fixed strategy yields 62.5% in both blocks but it uses information equally from both target and distractor; thus, it fails the second criterion. To test this, we used a paired Mann-Whitney U-test to compare whether the action on each trial was correct for the target versus correct for the distractor. In practice, none of the rats actually adopted such a hypothetical strategy: although some sessions failed the first criterion and were discarded (p > 0.05 for 4/55 sessions), no sessions failed the second criterion (p < .005 for all sessions). Therefore the first criterion (performance above 50% in both blocks) is actually the most relevant, and we mark the chance level on the plot as 50%.

**Construction and implantation of recording and stimulation microdrives**

We constructed tetrodes by cutting lengths of 12.5 micron nichrome wire coated with partially annealed polyimide insulation (Kantcl Palm Coast), twisting them, and heating with a heat gun until the 4 individual strands melted together. The tetrodes were then routed through polyimide guide tubes and glued to the moveable plastic tab within a potentiometer. We pinned the individual wires using gold pins (Neuralynx) into a custom printed circuit board (custompcb.com, beta-layout.com) that we designed. To reduce the Johnson-Nyquist noise at the electrochemical interface, which scales with the square root of impedance, electrode wires were gold-plated to 0.3 megaohm before implantation. For animals with dual implants, we built separate drives to target A1 and PFC. These were connected with a custom-designed adapter to a 32-channel preamp/headstage (Triangle BioSystems International).

Standard surgical techniques were used (Supp. Info.). Craniotomies were performed directly dorsal to the target areas (A1: 5.25 mm posterior and 6.5 mm left from bregma; prelimbic (PL) region of mPFC: 3.0 mm anterior and 1.0 mm left from bregma) and the tetrodes subsequently lowered downward into the target areas before recording.

In subject Z1 we implanted a pair of stimulating electrodes (FHC Inc., Model MX211FP-CR1, platinum-iridium, 240 um inter-electrode spacing, 400 kilohm impedance, rounded extra blunt taper) in each hemisphere centered near the dorsal portion of the prelimbic region, for a total of four electrodes. In subjects Z2 and Z3 we
implanted an array of three electrodes (MicroProbes microelectrode array, 70/30% PtIr, 100 kilohm target impedance) in each hemisphere spanning the anterior-posterior and dorsal-ventral extent of the prelimbic region, for a total of six electrodes, each 1.4-1.5 mm from its nearest neighbor. Using an isolated pulse stimulator (A-M Systems Inc., Model 2100), we injected a train of 1 ms current pulses at 10 Hz into mPFC on a subset of trials (“disruption trials”) and only during the center-poke hold and stimulus presentation. Stimulation was bipolar, first a 500 us positive and then a 500 us negative pulse of equal amplitude. Two rats (Z1 and Z3) were also implanted with recording drives in auditory cortex. Stimulation was always delivered in parallel to all channels at once; a stainless steel, cranial ground screw provided the return path. Typically this ground screw was distinct from any recording ground or reference screws.

We began with a very low current, around 10uA per electrode, which was typically too low to produce any behavioral effect. We wanted to use a minimal perturbation to ensure that the effects were as localized as possible in both time and space, and so we used pilot sessions to increase the amount of current until performance on the task became moderately impaired. During the testing sessions which we report in the main text, the mean currents used were 37 uA, 41 uA, and 23 uA per electrode for Z1, Z2, and Z3 respectively. See Supp Info for further details and comparison with other studies.

**Recording and signal processing**

The electrodes were lowered by approximately 100-200 microns before most recording sessions by turning the potentiometer’s screw. Before recording, we waited 30 minutes to allow the tetrodes to fully adjust. Broadband data were acquired at 30KHz and digitized and stored using a neural signal processor from Blackrock Microsystems. After the behavior, white noise bursts were presented passively to the animal to detect field and/or multi-unit auditory responses. Strong, low-latency auditory responses indicated that the electrodes were in A1 (in combination with the stereotactic coordinates used during implantation and, when possible, post-mortem histological reconstruction of electrode tracks). We only considered sessions in which we believed the electrodes to be in the correct brain regions.
We filtered the data offline to separate LFP (<200Hz) and spikes (>3 kHz). Butterworth, non-causal (temporally symmetric) filters were used to ensure that no phase distortion occurred. We used a detection threshold of 4.5 sigma (calculated using the more robust median absolute deviation) and a short window of 0.8 ms in order to minimize collisions between detected spikes. We extracted spike waveforms using our own contributions to the open-source OpenElectrophy software suite, reduced the dimensionality with principle component analysis, clustered with KlustaKwik, and manually reclustered as necessary with Klusters (Hazan et al., 2004) while blind to the experimental variables. Single units were identified based on the existence of a refractory period and minimal cluster overlap with other putative single units or noise.

We analyzed the data with Python and the modules numpy, scipy, scikits-learn, rpy2, statsmodels, and pandas, as well as custom-written data analysis code. Except where otherwise noted in the text, we observed consistent results across all subjects and therefore pooled the data (Figures S3D, E; S4D, E).

**Decoder analysis**

An ideal decoder was trained on the evoked rates, including baseline. We implemented this decoder using the LogisticRegression object in scikits-learn and assessed its performance by the number of trials on which the identity of the noise burst and/or warble was correctly predicted. Figure 6E shows the results on ensembles of simultaneously recorded neurons, but the results (no effect of block) were similar for individual neurons.
References


Figure Captions

Figure 1. Behavioral paradigm.

A) Left: a schematic of the behavioral arena with left (L), center (C), and right (R) ports (or nose-pokes), and left and right speakers. Right: timeline of each trial. The rat initiates a trial by nose-poking the center port, in the position shown on the left. After a hold period, an auditory stimulus plays in stereo. Following this, the rat may choose to go to the left port (blue arrow), go to the right port (red arrow), or do neither of those (a “nogo” response).

B) Task stimuli (left: description; right: spectrogram of the auditory waveform). On each trial, the rat hears one of four possible auditory stimulus pairs: LEFT+HIGH, RIGHT+HIGH, LEFT+LOW, or RIGHT+LOW. Each is a simultaneous combination of a broadband noise burst played from either the left or right speaker, and a low- or high-pitched warble. The warble is always played with equal intensity from both speakers.

C) Task rules. The session consists of alternating localization and pitch discrimination blocks of 80 trials each. Left: In localization blocks, the rat must go left for sounds containing LEFT and it must nogo for sounds containing RIGHT; the low- or high-pitched warble is an irrelevant distractor. Right: In pitch discrimination blocks, the rat must go right if the stimulus pair contains LOW and it must nogo if the stimulus pair contains HIGH; the localized noise burst is an irrelevant distractor. Good performance depends on selecting and responding to the target sound, not the distractor sound.
Figure 2. Trained rats select and respond to the target sound, not the distractor

A) Behavior performance during recording sessions. Each hash mark is the performance during localization (blue) and pitch discrimination (red) in a single recording session. Performance is well above chance (black dotted line, see Methods).

B) Distribution of behavioral responses to an example stimulus pair (RIGHT+LOW) over the course of an average session. We averaged across all sessions from a single rat (CR21A) and binned the trials into groups of 10 to smooth the traces. The x-axis shows both trial number and block type. The correct response to this stimulus pair is to go right during pitch discrimination and to nogo during localization (see Figure 1C). Each trace shows the probability that the rat will go right (red), nogo (gray), or go left (blue); black open squares mark the correct response for that block. The rat responds correctly most of the time, even though the required action changes abruptly at the block boundaries. Cue trials, during which this stimulus pair does not occur, begin each block and are shaded in cyan and pink throughout this figure.

C) Combined performance, similar to (B) but averaged over all sessions, rats, and stimuli. Correct responses (black trace) are the most common outcome. Performance is consistently high throughout, except immediately after a block change. The orange trace shows the probability of an “interference” trial (see text).

D) Analysis of performance immediately after block changes. All localization blocks from (C) are averaged together as are all pitch discrimination blocks. (In order to emphasize block transitions, the x-axis repeats itself after trial 160; the block structure is cyclical and so the cyan shaded areas are identical.) Immediately after the beginning of a new block (cyan and pink areas), performance decreases briefly but recovers within a few trials.
Figure 3. Pre-stimulus activity in mPFC encodes the selection rule

A) Left: An example mPFC single unit that fires more during the hold period for localization (blue bars throughout this figure) than for pitch discrimination (red bars); error bars SEM. Inset: Extracellular waveforms (mean plus or minus standard deviation) on each channel of the tetrode, duration 0.8 ms. The waveforms are colored red and blue based on the block in which they were recorded, but are almost entirely overlapping (purple). Right: peri-stimulus time histogram (PSTH) of the same unit, averaged over all correct trials from each block. The firing rate is significantly higher ($p < 0.001$) during the hold period (gray shading) for localization (mean 12.1 Hz, $n=483$ trials) vs. pitch discrimination (mean 7.2 Hz, $n=295$ trials). We assessed significance for all neurons with the Mann-Whitney U-test and controlled for multiple comparisons with the Benjamini-Hochberg false discovery rate.

B) Another example mPFC single unit, this one preferring pitch discrimination. This neuron’s firing rate is persistently elevated at all plotted timepoints. The hold period firing rate is significantly higher ($p < 0.001$) during pitch discrimination (mean 5.4 Hz) vs. localization (mean 2.7 Hz). Trial counts are the same as the simultaneously recorded unit in (A).

C) Stacked histogram of the ratio of hold period firing rate (pitch discrimination over localization) for all mPFC neurons. Red and blue bars are significantly modulated neurons. Quantifying this effect as the difference, rather than ratio, yielded similar results (Figures S3A, B).

D) Rule encoding during the hold period is diminished on interference trials. We averaged together the firing rates in the preferred and non-preferred blocks of each rule encoding neuron, after normalizing by subtracting the firing rate on correct trials in the non-preferred block. Error bars: SEM; orange bars: interference trials; white bars: correct trials. The population response on interference trials is significantly decreased during the preferred block and increased during the non-preferred block. Thus, the encoding of selection rule is
diminished on trials on which the rat may be selecting the wrong sound. Significance was assessed with a paired Mann-Whitney test (n=57 neurons), which is invariant to the subtractive normalization performed.
Figure 4. Pre-stimulus activity in A1 also encodes the selection rule

A) An example neuron recorded in primary auditory cortex (A1). This neuron responds significantly more (p < 0.001) during localization (8.0 Hz, n=312 trials; blue throughout this figure) than during pitch discrimination (4.8 Hz, n=253; red). Note the peak following stimulus onset, which was used to analyze the evoked response (Figure 6). Throughout this figure, we use the same conventions and statistical procedures as in Figure 3.

B) Another simultaneously recorded example A1 neuron that encoded the selection rule. This neuron significantly (p < 0.001) prefers pitch discrimination (10.1 Hz, n=312 trials) over localization (2.0 Hz, n=253).

C) Stacked histogram of the ratio of hold period firing rate (pitch discrimination over localization) for all A1 neurons. Quantifying this effect as the difference, rather than ratio, yielded similar results (Figures S4A, B).

D) Rule encoding during the hold period is inverted on interference trials for A1 neurons. Same conventions as Figure 3D, but the effect is stronger here. The population response on interference trials (orange bars) is significantly greater during the non-preferred block than during the preferred block (p < 0.05, n=16 neurons, paired Mann-Whitney U-test), opposite to the encoding on correct trials (white bars).
Figure 5. Within-trial timescale of the encoding of selection rule

A) PSTHs from example rule-encoding mPFC neurons in each block (blue: localization, red: pitch discrimination). Note that the timescale is much longer than in previous figures. Firing rates are smoothed with a 50ms Gaussian kernel, normalized to equal variance, and locked to stimulus onset at time 0 ms. The time interval containing the hold period during which the traces significantly diverge is shaded gray. Although these neurons were identified based on a difference in firing rate during the hold period, the traces often diverge for much longer than that. We observed a wide variety of timescales and dynamics in the block-specific anticipatory modulation. The first neuron effectively fires persistently more in one block. The third and fourth neurons demonstrate that the firing rate can either rise during the preferred block, or, less commonly, drop during the non-preferred block. The fifth neuron shows that the anticipatory effect can be limited to just the hold period alone.

B) Example neurons from A1, following the conventions of (A). Again, the neurons exhibit a wide variety of dynamics, from essentially persistent block-specific activation for over three seconds preceding the stimulus (first neuron), to very brief activation well under 1 second (last neuron). The third neuron shows increased baseline firing and increased stimulus-evoked firing (peak immediately after time zero) in the same block. This was typical of our dataset (see Figure 6).

C) Population timecourse: the curves represent the average response during the preferred block across all rule-encoding neurons in mPFC (purple) and A1 (orange). The firing rates of all rule-encoding neurons were normalized (mean: 0, variance: 1) and then averaged together. Only the response during the preferred block is shown. Traces are mean response (plus or minus SEM) across neurons. Thick mean trace: timepoints during which the population response significantly exceeds zero, the mean firing rate (p < 0.05, one-sample t-test across neurons). In both populations, the firing rate in the preferred block shows a gradual increase, peaking around the time of stimulus onset, and then decreases more quickly back to baseline. The PFC population increases its response earlier (first significantly activated 2.4 s before stimulus onset, n=76) than the A1
population (first significantly activated 0.78 s before stimulus onset, n=36), consistent with the hypothesized role of PFC as the source of top-down modulation.

D) Population time course, plotted separately for go and nogo trials. Left: peri-event time histograms (PETHs), locked to the post-stimulus exit from the center-port, from rule-encoding mPFC neurons during their preferred block. Right: same as left panel, but for A1 neurons. Each PETH is aligned to the post-stimulus exit from center-port. Trials are grouped according to correct go responses (green) and correct nogo responses (gray). As in (C), PETHs were normalized to unit variance and zero mean before averaging across neurons; only the response during the preferred block is shown. Trace thickness indicates SEM across neurons. On nogo trials, the firing rate remains elevated above baseline for at least several seconds, during which time the rat typically had already initiated the next trial. On go trials, the firing rate falls below baseline and remains there as the rat moves to the choice port and drinks a reward (which always required at least several seconds). To illustrate this last point, along the lower edge of the figure we also plot the distribution of latencies to the relevant trial events: stimulus onset (black), reward delivery (green, go trials only), and center-poke beginning the next trial (gray, nogo trials only; for go trials the beginning of the next trial would not be visible in this time range due to the time necessary to consume reward).
Figure 6. Limited evidence for modulation of stimulus-evoked activity

A) An example A1 neuron exhibiting a preference for some acoustic stimuli (LEFT+HIGH, LEFT+LOW) over others (RIGHT+HIGH, RIGHT+LOW), but no change in this tuning with block (localization: blue; pitch discrimination: red). Black triangle: stimulus onset; shaded area: response window for this neuron.

B) An example mPFC neuron that responds to the task stimuli with a low-latency response. Auditory responses were weaker in mPFC neurons compared with A1 neurons (Figure S6A-E).

C) For A1 neurons, increase in hold period activity during one block correlates with increased evoked response during that block. For each neuron, the change in evoked response (driven spikes in pitch discrimination vs. localization) is plotted against the change in hold period firing rate (anticipatory spikes in pitch discrimination vs. localization). The trend line (n=43 neurons, r=0.52) has a slope of 0.98, suggesting that much of the modulation of evoked strength is due to anticipatory modulation (example: Figure 4B).

D) Following the conventions of (C), but for auditory-responsive PFC neurons. Again, a change in baseline activity correlates closely with a change in evoked activity (n=17 neurons, slope=1.46, r=0.85).

E) No evidence for tuning changes that increase the decodability of the target sound. The identity of the noise burst (LEFT or RIGHT) or the warble (LOW or HIGH) can be decoded from the trial-by-trial responses of simultaneously recorded ensembles of auditory-responsive cells in either A1 or PFC. It can be decoded significantly better (p < 0.001) from A1 cells (n=22 ensembles of 57 neurons total) than from mPFC cells (n=13 ensembles of 25 neurons total), but it cannot be decoded significantly better during either block. The chance decoding level, attainable by a neuron with no information about the stimulus, is 0.5. The mean and SEM over the ensembles is shown. Significance was assessed with a 3-way ANOVA on brain region, target sound, and block.
Figure 7. Disruption of mPFC robustly impairs performance on the task

A) Electrical disruption of mPFC significantly impacted task performance during localization trials (left panel) and/or pitch discrimination trials (right panel) in most sessions. Each point represents the performance within a single session on control (x-coordinate) vs. disruption trials (y-coordinate). Plus symbols represent sessions during which the performance was significantly impaired (p < 0.05, Fisher’s exact test). Throughout this figure, different colors represent different rats (red: Z1, yellow: Z2, green: Z3).

B) Example session from each rat. Performance is shown for each trial type (go and nogo in each block). Solid bars represent control trials; open bars represent disruption trials. Error bars: 95% confidence intervals using Pearson-Klopper binomial fit. Asterisks indicate trial types for which electrical disruption significantly impairs performance (Fisher’s exact test). The effect is robust within each example session, but variable across rats. See Supp. Info. for the data from all sessions.

C) Impairment on each trial type for each rat, across sessions. Error bars: SEM across sessions. All rats showed a significant impairment on nogo trials in one block or the other (p < 0.05, binomial test on the number of sessions demonstrating impairment). One rat (Z2) also showed a significant impairment on localization go trials.
Figure 8. A simulated network model demonstrates how anticipatory modulation could solve the stimulus selection problem.

A) Network connectivity. A1 consists of a population of N neurons, each with random tuning for the four task stimuli and subject to additive Gaussian noise. Red and blue neurons in A1 are differentially activated in one block or the other, based on an excitatory “task signal” projection, hypothetically originating in PFC. Each subpopulation in A1 connects to a set of premotor command neurons encoding the possible responses in that block. The model’s choice is determined by which command neuron is the most active via a winner-take-all mechanism. The weights W1 and W2 are constrained to be excitatory (non-negative) and are separately optimized during an initial supervised training phase, then fixed.

B) Performance of the model for N=320 neurons on task 1 (left panel) and task 2 (right panel). We tested a range of values for the sensory signal-to-noise ratio (SNR), defined as the ratio of the tuning for sensory stimuli to the strength of the additive Gaussian noise in each A1 neuron. We plot the probability of a correct choice versus the strength of the task signal. For the highest SNR of 0.25 (darkest trace), the model produces 100% correct responses for virtually any positive task switch signal. (Negative task signals correspond to activating the subpopulation corresponding to the incorrect block.) As the task signal increases in strength, the sensory input is eventually drowned out and the model’s performance falls to chance (50%). This problem is especially pronounced at the lowest SNR, near 0.015. However, larger networks can still perform well at such SNRs (Supp. Info.).
Figure 1

B  | Stimulus Pair       | Abbrev. Name  |
---|---------------------|---------------|
left noise burst + high warble | LEFT + HIGH |
right noise burst + high warble | RIGHT + HIGH |
left noise burst + low warble | LEFT + LOW |
right noise burst + low warble | RIGHT + LOW |

C  | Localization |
---|--------------|
LEFT + HIGH | RIGHT + HIGH |
LEFT + LOW  | RIGHT + LOW  |

Correct Response
- go left
- nogo
- go right
Throughout: localization, pitch discrimination

Figure 3
Throughout: localization

Figure 4

A

Spike rate (Hz)

Block number

B

Spike rate (Hz)

Block number

C

13/99 prefer localization

23/99 prefer pitch disc.

distribution of neurons

number of neurons

ratio of spike rate (fold)

D

population response (Hz)

hit

inter.

hit

inter.

non-preferred block

hit preferred block

hit preferred block

inter.

**
Figure 5

A. Example PFC neurons

B. Example A1 neurons

C. Normalized population response

D. Normalized firing rate

Time from stimulus onset (s)

Time from exiting center poke (s)

Pitch discrimination

Localization

PFC

A1

Stim onset

Reward

Next trial

NOGO

GO
Figure 6 throughout: localization | pitch discrimination

A

LEFT+HIGH | LEFT+LOW | RIGHT+HIGH | RIGHT+LOW

25 Hz
50 ms

B

LEFT+HIGH | LEFT+LOW | RIGHT+HIGH | RIGHT+LOW

10 Hz
50 ms

C

Evoked change (PD vs L, Hz) vs baseline change (PD vs L, Hz)

D

Evoked change (PD vs L, Hz) vs baseline change (PD vs L, Hz)

E

Decodability of noise burst or warble from neuronal firing rates

- noise burst
- warble

A1 | A1 | PFC | PFC

- pitch discrimination
- localization

ns

***
Figure 7

A

localization

Z1 n=6
Z2 n=8
Z3 n=8

pitch discrimination

Z1 n=6
Z2 n=8
Z3 n=8

control performance

disruption performance

control performance

B

performance

Z1

***

Z2

***

Z3

***

local.

pitch disc.

go

nogo

go

nogo

local.

pitch disc.

go

nogo

local.

pitch disc.

go

nogo

local.

pitch disc.

go

nogo

C

impairment

* * *
Figure 8

A1 trained for task 1, then fixed

A1 output projections

W1 trained for task 1, then fixed

goose

nogo

left+high

right+high

left+low

right+low

premotor command neurons

W2 trained for task 2, then fixed

goose

nogo

go right

take all

winner

choice

B

Sensory SNR

Sensory SNR

prob(correct for task 1)

prob(correct for task 2)

strength of task 1 signal

strength of task 2 signal

2^{-2}

2^{-3}

2^{-4}

2^{-5}

2^{-6}
Inventory of Supplemental Information
Rodgers and DeWeese [manuscript]

Supplemental Data

Figure S2, relating to Figure 2: Behavioral performance 1
Figure S3, relating to Figure 3: Notes on mPFC anticipatory effect 3
Figure S4, relating to Figure 4: Notes on A1 anticipatory effect 8
Figure S5, relating to Figure 5: More information on timecourse 12
Figure S6, relating to Figure 6: More information on evoked activity 14
Figure S7, relating to Figure 7: More information on electrical disruption of mPFC 17
Figure S8, relating to Figure 8: Behavior of the model over parameter space 18

Supplementary Experimental Procedures

Behavioral parameters 19
Surgical implantation 19
Analysis of possible confounds: waveform variation and firing rate drift 19
Power analysis 20
Calculation of evoked response 20
Changes in evoked response not explained by baseline 20
Disruption of mPFC by electrical stimulation 21
Simulated network model 22
Rule-encoding is not dependent on the motion history of the animal 23
Video analysis of preparatory head positioning 24
Verification of ANOVA 25
Posture equalization by trial selection 27
Acknowledgements 28
Supplemental Data

Each figure is linked to the main figure with the same number. Not all main figures have associated supplementary figures.

Figure S2. Behavioral performance, related to Figure 2

A) Performance of each rat in greater detail, with go and nogo trials separately considered in each block. Rats generally did better on go (first and third columns) than on nogo trials (second and fourth columns). Some rats did better on localization (first and second columns) than on pitch discrimination (third and fourth columns). Error bars show SEM across sessions.

B) Performance (fraction of correct responses) of one rat that performed a slightly modified “catch trial” task on the last day of recordings. This task was designed to probe whether the rats learn to respond to a unified stimulus pair, or whether they learn to respond just to the target sound regardless of the identity of the distractor. On a small proportion (15%) of trials, we replaced the distractor with a neutral sound that the rat had never heard on any prior training session or behavioral testing session. For example, on catch trials during localization the rat heard the same target as before (LEFT or RIGHT) with a novel mid-range warble with no behavioral relevance. If the rats had memorized each of the four possible stimulus pairs and were unable to generalize properly to cases with an intact target sound combined with a novel distractor, they should have performed at chance on these novel stimulus pairs. The performance on catch trials (red) and standard trials (white) for each trial type is shown, with 95% confidence intervals (Pearson-Klopper binomial fit). The rats perform just as well on catch trials as on standard trials (Fisher’s Exact Test on the trial outcomes, $p > 0.05$ in all cases). This suggests that the rats are selecting the target stimulus, not memorizing a fixed set of four stimulus pairs.

C) Same as panel B, but for a different rat.
FIGURE S2
**Figure S3. PFC hold period, related to Figure 3**

A) Stacked histogram of the difference, rather than ratio, of hold period firing rate between blocks for all mPFC neurons.

B) Alternative presentation of the hold period effect across mPFC neurons. The hold period firing rate in each block is shown as the x- and y-coordinate of each point (red and blue: significant block preference; gray: not significant). Note the logarithmic scaling, to avoid crowding the points with low firing rate. Error bars are 95% confidence intervals obtained by bootstrapping and were truncated at the edge of the plot. Significance was assessed with a Mann-Whitney U-test as described in the text.

C) Analysis of the hold period effect in mPFC on various types of correct and error trials. The trials are grouped by the meaning of the target sound, the rat’s response, and the meaning of the distractor sound. Neurons are grouped by their preferred block. White bars represent correct trials, on which the rat’s response matches the target sound. The gray bar represents “go-on-nogo” error trials when the target required a nogo response but the rat went to the choice port anyway. (The opposite error, nogo-on-go, was too rare to include in this analysis. We only analyzed neurons from sessions with at least 3 trials of each type, where type refers to each bar in this panel.) The orange bar represents “interference” trials on which the rat heard a distractor sound cuing a go response and went to the choice port associated with that distractor (WP, or wrong port). To aid in visualization, firing rates were normalized by subtracting the mean response on correct trials during the non-preferred block, and then averaged across neurons. There is no significant difference in the hold period activity between correct go, correct nogo, and incorrect go-on-nogo trials. If the hold period activity actually encoded a simple motor plan to go to the choice port regardless of the upcoming stimulus, then it should have been lower on correct nogo trials when the rat did not perform this action. Thus, if the hold period activity represents a motor plan, it must be subject to change (“countermanding”) after the stimulus. However there is a significant difference between interference trials and all other trial types in that block. That is, when the rat gives the response that would be appropriate in the other block, the anticipatory activity is higher in the non-preferred block and lower in the preferred block — i.e., the block modulation is attenuated, tending toward reversal of the effect. Significance between each pair of bars was assessed with a paired Mann-Whitney U-test across neurons, which is invariant under the subtractive normalization performed, and the p-values were Bonferroni corrected.

D) Histogram of the number of neurons preferring pitch discrimination (red), localization (blue), and neither (black). In some rats (marked N/A), no mPFC neurons were recorded.

E) Proportion of rule-encoding neurons in mPFC is consistent across rats. The data from (D) are expressed here as a percentage of total neurons. No percentages are plotted for rats with fewer than 8 neurons total recorded in mPFC.

F) On average, firing rates in both blocks are bidirectionally modulated, compared with the spontaneous rate. We defined the spontaneous rate as the average rate during epochs more than 2 s from the nearest stimulus onset. Localization-preferring neurons (left) fire significantly more than spontaneous in localization and significantly less than spontaneous in pitch discrimination. A similar statement holds for pitch discrimination-preferring neurons (center). Neurons that do not prefer either block (right) — that is, neurons for which the pre-stimulus firing rate does not significantly differ between blocks — tend to be suppressed (relative to their spontaneous rates) in both blocks, though this trend was not statistically significant during pitch discrimination (p = 0.06). The bars show the average and SEM across neurons of the hold period firing rates minus the spontaneous firing rate for that neuron. We assessed significance with a paired Mann-Whitney test.

G) No evidence for a correlation between the pre-stimulus firing rate of PFC neurons and the animal's reaction time, defined as the time between stimulus onset and withdrawal from center port. We plot here the distribution of correlation coefficients obtained in both blocks, with red representing significantly correlated neurons (p < 0.05 after correction with the false discovery rate). Only a small minority of neurons showed a significant
correlation, and the mean of the full distribution was not significantly different from zero (p > 0.05, one-sample t-test). Similar results were obtained when considering only rule-encoding neurons — those with a significantly increased pre-stimulus firing rate in one block or the other.

H) Similar to (G) except that here we are correlating the anticipatory firing rate with the “motion time,” which we defined as the time necessary for the rat to move to the choice port on successful go trials. Again, only a small minority of neurons show an individually significant correlation and the population distribution is not significantly different from zero. Similar results were obtained when considering only rule-encoding neurons.

I) The azimuthal (left/right) angle of the rat’s head during center poke, relative to the behavior box, differs by approximately 30 degrees between blocks. Red: pitch discrimination. Blue: localization. The mean head angle over all trials is defined here as zero degrees. These data are from an example session but all analyzed sessions yielded similar results. This preparatory motor activity is presumably an adaptive behavioral strategy in response to the fact that the choice port differs between blocks. See “Video analysis of preparatory head positioning” in the Supp. Exp. Proc. for further detail on panels I-L.

J) Example PFC neuron recorded during the session shown in (I). Each point shows the square root of the number of spikes fired and head angle on a single trial. (Throughout this figure, we use the square root transformation to stabilize the variance of the counts and thus to ensure homogeneity of variance independent of the mean to facilitate our regression analysis.) Across all trials, these variable are significantly correlated (black trend line). However, this correlation between firing rate and head angle is almost entirely explained by block type. Within each block, there is no such correlation — the red and blue trend lines are not statistically significantly different from horizontal.

K) Fraction of explainable variance (FEV) in the spike counts (again square root-normalized) that the least-squares linear fit attributes to block (orange) and head angle (black), individually for each rule-encoding PFC neuron during the analyzed video sessions. The magenta horizontal line indicates 50%. The bars are sorted in order of increasing head angle FEV. Most bars are mostly orange, indicating that block is the major explanatory factor in the spike count. Only three out of 16 bars are more than 50% black, corresponding to neurons whose firing rate was mostly explained by head angle.

L) Summary plot of the data in panel (K). The distribution of FEV across neurons is plotted for each factor (head angle and block). Individual crosses represent individual rule-encoding PFC neurons. The red line shows the median; the gray box outlines the inter-quartile range. Across the population, most of the variance (median: 94.1%) is explained by block; the small amount that remains is explained by head angle (median: 5.9%).

M) Rule-encoding in mPFC neurons is broadly similar to that of the full data set when only “posture-equalized” trials are considered. Here, we analyze only rule-encoding neurons (i.e., those with a significant preference for one block) recorded during the three sessions for which we analyzed the video and obtained head angle measurements. Throughout this panel, the x-axis shows the magnitude of rule encoding. We quantified rule encoding in panels M, N, and O of this figure as the base-10 logarithm of the ratio of the firing rate in pitch discrimination over localization. Top: histogram of rule encoding over neurons, when considering only posture-equalized trials. Middle: same as above, but for the rest of the trials. Bottom: distribution of rule encoding on the posture-equalized trials after shuffling the block labels. Rule encoding is similarly distributed on both posture-equalized (top) and other trials (middle). Rule encoding is significantly (p < 0.01; shuffle test) stronger on posture-equalized trials (σ of 0.47) than would be obtained by random chance (mean σ of 0.10), as in the bottom row. See “Posture equalization by trial selection” in Supp. Exp. Proc. for additional details.

N) Mean magnitude of rule-encoding in mPFC cells is not significantly altered on the posture-equalized trials. The mean and S.E.M. of the magnitude of rule encoding is shown for both localization- and pitch discrimination-preerring cells (n=8 for both groups), on posture-equalized trials (blue bars) or all other trials
There is no significant (p > 0.05, paired Mann-Whitney U-test) difference in the mean magnitude of rule encoding, though there is a non-significant trend toward weaker rule encoding on posture-equalized trials.

O) Posture-equalization results for individual PFC cells (n=16). The rule encoding of each cell is plotted for posture-equalized trials (left) and all other trials (right) and connected with a line. The red line indicates a rule encoding of zero: no difference between blocks. In some cells the rule encoding is weaker (closer to zero) on posture-equalized trials. If head angle were the only determinant of the difference in firing rate, then we would expect rule encoding on posture-equalized trials to be around zero, randomly greater in localization or pitch discrimination with 50% probability. Instead, we see that in only one out of 16 mPFC cells does the preferred block reverse on the posture-equalized trials, versus the rest of the trials. If reversals were random, then the probability of observing such a small ratio of reversals is 0.0005 (two-tailed binomial test).

P) Rule encoding persisted throughout reward consumption in some neurons. This panel shows example PSTHs during localization (blue trace; line thickness indicates SEM over trials) and pitch discrimination (red trace) of two example mPFC neurons that maintain rule-encoding throughout reward consumption. The traces are locked to exiting the center port at time zero. Along the bottom of the panel, distributions of the latencies to stimulus onset (black) and entering reward port (green) are shown. Only go trials are included. Reward consumption began upon entry into the reward port and continued for several seconds.

Q) Simulated posture-encoding neurons are correctly detected by the multiple-regression/ANOVA approach used to analyze the real data. For each rule-encoding mPFC neuron in (K) and (L), we designed a model neuron with a firing rate linearly related to the head angle, but not the block, on each trial. We added Gaussian noise to its firing rate to match the variance of the actual counts. Finally we analyzed each model neuron using the same multiple-regression/ANOVA approach that we used in (K) and (L). Here, the model neurons are correctly identified as primarily encoding head angle: the median FEV of head angle is 0.867 and it is above 0.5 for every neuron. Compare with the results in (K) and (L) in which few neurons were identified as primarily encoding head angle (median FEV of head angle: 0.059). This demonstrates that the analysis is capable of detecting postural encoding where it exists. See “Verification of ANOVA on simulated posture-encoding neurons” in Supp. Exp. Proc. for further details.

R) Estimated FEV (y-axis) of HeadAngle versus its true FEV (x-axis), using the empirical distributions for head angle and block identity and a simulated model neuron that linearly combines these two variables. Error bars represent the standard deviation over 150 runs. Black unity line shows perfect estimate. The estimate closely tracks the true value. See “Verification of ANOVA with correlated regressors” in Supp. Exp. Proc. for further details on panels R-T.

S) Same as (Q), except that head angle and block identity are drawn from a random normal and binomial distribution, respectively, subject to the constraint that the correlation between these two values is 0.72. The estimated FEV still closely tracks the true FEV.

T) Same as (Q), except that the output of the model neuron is converted into a discrete, quantized representation more closely resembling spiking data.
FIGURE S3 (continued on next page)
FIGURE S3 (continued from previous page)
Figure S4. A1 hold period, related to Figure 4
A, B) Same as Figure S3A,B, but for A1 neurons instead of mPFC neurons.

C) Same as Figure S3C (correct trials: white; go-on-nogo errors: gray; interference trials: orange), but for A1 neurons instead of mPFC neurons. The effects are similar to those for mPFC, but note that the effect on interference trials is stronger. For A1 neurons, the direction of hold period modulation is significantly reversed — the firing rate is higher during such trials in the non-preferred block than in the preferred block. However, there is no difference between the other trial types, regardless of whether the rat performed a go or nogo response.

D) Histogram of the number of neurons preferring pitch discrimination (red), localization (blue), and neither (black). In one rat, no A1 neurons were recorded (marked N/A).

E) Proportion of rule-encoding neurons in A1 is consistent across rats. The data from (D) are here expressed as a percentage of total neurons. No percentages are plotted for rats with fewer than 8 neurons total recorded in A1.

F) On average, firing rates in both blocks are bidirectionally modulated, in comparison to the spontaneous rate. We defined the spontaneous rate as the average rate during epochs more than 2 s from the nearest stimulus onset. Localization-prefering neurons (left) fire significantly more than spontaneous in localization and significantly less than spontaneous in pitch discrimination. A similar statement holds for pitch discrimination-prefering neurons (center). Neurons that do not prefer either block (right) — neurons for which the pre-stimulus firing rate does not significantly differ between blocks — showed no change in their firing rates relative to their spontaneous rates. The bars show the average and SEM across neurons of the hold period firing rates minus the spontaneous firing rate for that neuron. We assessed significance with a paired Mann-Whitney test.

G) No evidence for correlation between pre-stimulus firing rate of A1 neurons and the animal’s reaction time, defined here as the time between stimulus onset and withdrawal from center port. We plot here the distribution of correlation coefficients obtained in both blocks, with red representing significantly correlated neurons (p < 0.05 after correction for multiple comparisons with the false discovery rate). Only a small minority of neurons showed a significant correlation, and the mean of the full distribution is not significantly different from zero (p > 0.05, one-sample t-test). Similar results were obtained when considering only rule-encoding neurons — those with a significantly increased pre-stimulus firing rate in one block or the other.

H) Similar to (G), but here we correlate the anticipatory firing rate with the “motion time” — the time necessary for the rat to move to the choice port on successful go trials. Again, only a small minority of neurons show an individually significant correlation and the population distribution is not significantly different from zero. Similar results were obtained when considering only rule-encoding neurons.

I) The azimuthal (left/right) angle of the rat’s head during center poke, relative to the behavior box, differs by approximately 30 degrees between blocks. Red: pitch discrimination. Blue: localization. The mean head angle over all trials is defined here as zero degrees. These data are from an example session but all analyzed sessions yielded similar results. This preparatory motor activity is presumably an adaptive behavioral strategy in response to the fact that the choice port differs between blocks. See “Video analysis of preparatory head positioning” in the Supp. Exp. Proc. for further detail on panels I-L.

J) Example A1 neuron recorded during the session shown in (I). Each point shows the square root of the number of spikes fired and head angle on a single trial. (Throughout this figure, we use the square root transformation to stabilize the variance of the counts and thus to ensure homogeneity of variance independent of the mean to facilitate our regression analysis.) Across all trials, these variable are significantly correlated (black trend line). However, this correlation between firing rate and head angle is almost entirely explained by
block type. Within each block, there is no such correlation — the red and blue trend lines are not significantly different from horizontal.

K) Fraction of explainable variance (FEV) in the spike counts (again square root-normalized) that the least-squares linear fit attributes to block (orange) and head angle (black), individually for each rule-encoding A1 neuron during the analyzed video sessions. The magenta horizontal line indicates 50%. The bars are sorted in order of increasing head angle FEV. Nearly every bar is mostly orange, indicating that block is the major explanatory factor for the spike count across the population. Only one bar is more than 50% black, corresponding to a neuron whose firing rate was mostly explained by head angle.

L) Summary plot of the data in panel (K). The distribution of FEV across neurons is plotted for each factor (head angle and block). Individual crosses represent individual rule-encoding A1 neurons. The red line shows the median; the gray box outlines the inter-quartile range. Across the population, most of the variance (median: 83.0%) is explained by block; the small amount that remains is explained by head angle (median: 17.0%).

M) Rule-encoding in A1 neurons is broadly similar to that of the full data set when only “posture-equalized” trials are considered. Here, we analyze only rule-encoding neurons (i.e., those with a significant preference for one block) recorded during the three sessions for which we analyzed the video and obtained head angle measurements. Throughout this panel, the x-axis shows the magnitude of rule encoding. We quantified rule encoding in panels M, N, and O of this figure as the base-10 logarithm of the ratio of the firing rate in pitch discrimination over localization. Top: histogram of rule encoding over neurons, when considering only posture-equalized trials. Middle: same, but for the rest of the trials. Bottom: distribution of rule encoding on the posture-equalized trials after shuffling the block labels. Rule encoding is similarly distributed on posture-equalized trials (top) as it is on other trials (middle). Rule encoding is significantly (p < 0.01; shuffle test) stronger on posture-equalized trials (σ of 0.32) than would be obtained by random chance (mean σ of 0.09), as in the bottom row. See “Posture equalization by trial selection” in Supp. Exp. Proc. for additional details.

N) Mean magnitude of rule-encoding in A1 cells is not significantly altered on the posture-equalized trials. The mean and S.E.M. of the magnitude of rule encoding is shown for both localization- and pitch discrimination-preffering cells (n=4 for both groups), on posture-equalized trials (blue bars) or all other trials (orange). There is no statistically significant (p > 0.05, paired Mann-Whitney U-test) difference in the mean magnitude of rule encoding, though there is a non-significant trend toward weaker rule encoding on posture-equalized trials.

O) Posture-equalization results for individual A1 cells (n=8). The rule encoding of each cell is plotted for posture-equalized trials (left) and all other trials (right) and connected with a line. The red line shows a rule encoding of zero: no difference between blocks. In some cells the rule encoding is weaker (closer to zero) on posture-equalized trials. If head angle were the only determinant of the difference in firing rate, then we would expect rule encoding on posture-equalized trials to be close to zero, randomly greater in localization or pitch discrimination with 50% probability. Instead, we see that in none of the 8 A1 cells does the preferred block reverse on the posture-equalized trials, as compared with the rest of the trials. If reversals were random, then the probability of observing such a small ratio of reversals is 0.008 (two-tailed binomial test).

P) Rule encoding persisted throughout reward consumption in some neurons. This panel shows example PSTHs during localization (blue trace with thickness SEM over trials) and pitch discrimination (red trace) of two example A1 neurons that maintain rule-encoding throughout reward consumption. The traces are time-locked to exiting the center port at time zero. Along the bottom of the panel, distributions of the latencies to stimulus onset (black) and entering reward port (green) are shown. Only go trials are included. Reward consumption began upon entry into the reward port and continued for several seconds.

Q) Simulated posture-encoding neurons are correctly detected by the multiple-regression/ANOVA approach used to analyze the real data. For each rule-encoding A1 neuron in (K) and (L), we designed a model neuron with a firing rate linearly related to the head angle, but not the block, on each trial. We added Gaussian noise
to its firing rate to match the variance of the actual counts. Finally we analyzed each model neuron using the same multiple-regression/ANOVA approach that we used in (K) and (L). Here, the model neurons are correctly identified as primarily encoding head angle: the median FEV of head angle is 0.872 and it is above 0.5 for every neuron. Compare with the results in (K) and (L) in which few neurons were identified as primarily encoding head angle (median FEV of head angle: 0.170). This demonstrates that the analysis is capable of detecting postural encoding, where it exists. See “Verification of ANOVA on simulated posture-encoding neurons” in Supp. Exp. Proc. for further details.
FIGURE S4 (continued from previous page)
Figure S5. More information on timecourse, related to Figure 5

A) Similar to Figure 5C in the main text but with each trace time-locked to entry into the center port that initiated the trial, rather than the subsequent stimulus onset. This shows the population time course during the preferred block, averaged over all rule-encoding neurons in each region (purple: mPFC; orange: A1). Throughout this figure, firing rates were first normalized to zero mean and unit variance and then averaged over neurons; the thickness of the trace represents SEM over neurons. This demonstrates more clearly that the increase in anticipatory activity definitely precedes the center-poke entry.

B) Population time course during the preferred block for rule-encoding neurons in mPFC (left) and A1 (right). In this panel, we include only trials following a successful nogo response. Along the lower edge of the figure we also plot the distribution of latencies to the end of the previous trial, defined as the withdrawal from the center port following the nogo stimulus onset. The increased activity persists between the trials, even though the rat is typically remaining motionless near the center port during this time. This demonstrates that the anticipatory effect is not due to the rat’s motion toward the center port.

C) Similar to Figure 5D in the main text, but instead showing activity in the non-preferred block (green) in addition to activity during the preferred block (purple); also, only correct nogo trials are included to demonstrate the persistence of the effect. The population time course is locked to exit from center-port after the stimulus onset. Before exiting the center port, average activity in the non-preferred block is suppressed below baseline; average activity in the preferred block is increased above baseline. This remains true for at least several seconds after the rat exits the center port, during which time the rat typically had already initiated the next trial. To illustrate this last point, along the lower edge of the figure we also plot the distribution of latencies to the relevant trial events: stimulus onset (black) and the center-poke initiating the next trial (gray).

D) Similar to panel (C), except that here we include only correct go trials. After exiting the center port, the firing rate during the preferred block (again averaged over neurons) falls below baseline and remains there as the rat moves to the choice port and drinks a reward (which always required at least several seconds). In contrast, during the non-preferred block A1 neurons are actually activated above baseline during this period, thus inverting the usual rule encoding. Along the lower edge of the figure we also plot the distribution of latencies to the relevant trial events: stimulus onset (black) and reward delivery (green). Finally, we note that these results depict the population average; individual neurons displayed a wide variety of dynamics (Figure 4).

E) Similar to (C) and (D), but here we separately plot the population activity during localization (blue) and pitch discrimination (red) blocks from neurons preferring localization (left column) and pitch discrimination (right column) on correct nogo trials (top two rows) and correct go trials (bottom two rows). The dynamics of the pitch discrimination-prefering and localization-prefering neurons are quite similar (compare left and right columns), other than the fact that the red and blue traces are roughly reversed (by definition of preferred block). Along the lower edge of the figure we also plot the distributions of latencies to stimulus onset (black), reward delivery (green, go trials only), and center-poke beginning next trial (gray, nogo trials only).
FIGURE S5
Figure S6. More information on evoked activity, related to Figure 6
A) PSTH of a typical A1 neuron, to illustrate the notion of "evoked response strength." This neuron’s response strength is near the median of the A1 population. All stimuli and trials are included in this PSTH. Note that the response to the onset of the sound is rapid, short-latency, tightly stimulus-locked, and brief. The onset window is shaded. This was defined as the continuous set of time bins post-onset over which the firing rate was significantly greater than the rate preceding the stimulus. Some mPFC units also showed similar (though weaker) stimulus-driven responses.

B) An example A1 neuron showing one of the strongest and most sustained recorded responses.

C) Distribution of onset response strengths across n=49 auditory-responsive A1 neurons (blue) and n=31 auditory-responsive PFC neurons (orange). The response strength is expressed as the average number of additional spikes (over baseline) during the onset response window for that neuron. All trials and stimuli are included, and the baseline rate is calculated over the 50 ms preceding the stimulus onset. The responses are significantly stronger in A1 neurons (median: 0.11 spikes) than in PFC neurons (median: 0.02 spikes), p < 0.05, unpaired Mann-Whitney U-test. Note the long tail of the distribution: a small subpopulation fires much more strongly than the median.

D) Alternative presentation of onset response strength. The data are the same as in (C), but expressed here as percentage of baseline firing rate. Because the response window is so brief, a small number of additional spikes over baseline typically represents a large (many-fold) increase in rate. Again the responses are stronger in A1 neurons (median: 209% of baseline) than in PFC neurons (median: 171% of baseline), p < 0.001, unpaired Mann-Whitney U-test.

E) Distribution of latencies to center of onset response window across the same populations as (C) and (D). The PFC latencies are significantly longer (median: 19.75ms) than the A1 latencies (median: 16.75ms). p < 0.05, unpaired Mann-Whitney U-test.

F) No correlation between stimulus tuning and the change in pre-stimulus rate for auditory-responsive A1 neurons. We calculated the strength of the tuning for the noise burst (LEFT versus RIGHT) or the warble (LOW versus HIGH) using the difference in firing rate to those stimuli when they were presented on cue trials. Similar to Figure 6E in the main text, we quantified tuning using the performance of an ideal decoder. This metric has a value of 0.5 for identical responses, and 1.0 for perfectly discriminable responses. We found no correlation between how well the neurons were tuned for either stimulus and the change in pre-stimulus rate across blocks. Although these results were obtained using the cue trials to measure tuning, similar results were obtained when we calculated the tuning using the responses to the stimuli containing distractors (data not shown).

G) Same as (F), but for auditory-responsive PFC neurons. Note the difference in the scale of the x-axis vs (F), reflecting the fact that PFC neurons tend to be more poorly tuned for the stimuli than A1 neurons. Again there is no significant correlation, though there is a weak, non-significant trend in the following direction: neurons that are well-tuned for the warble tend to show increased anticipatory firing in pitch discrimination; similarly, neurons that are well-tuned for the noise burst tend to show increased anticipatory firing during localization.

H) Across the population of auditory-responsive neurons in both brain regions, all four stimulus pairs elicit equally small and non-significant changes across blocks in their evoked firing rate. Each box and whiskers plot shows the distribution across neurons of the difference in evoked firing rate between pitch discrimination and localization for each stimulus pair. (We used square root as a normalizing transform but this did not affect the results.) The box’s top and bottom show the inter-quartile range (IQR); the tapered area of each box shows the 95% bootstrapped confidence intervals on the median. Pluses indicate outliers, defined as more than 1.25 times the IQR from the median; these points were still included in the analysis. We also repeated the analysis after subtracting the pre-stimulus rate from the evoked rate in both blocks (labeled “baseline-subtracted”). A
one-way Kruskal-Wallis test revealed no significant difference between the stimulus pairs. We also assessed whether the change to each individual stimulus pair was different from zero using the Wilcoxon signed-rank test, correcting for multiple comparisons within each group (e.g., each group of four boxplots) using the false discovery rate. We found no significant difference from zero for any individual stimulus pair, indicating that there was no consistent trend toward increased evoked firing rates in either block for any stimulus pair.

I) Similar to (H), but here we took the absolute value of all of the data. We reasoned that some stimulus pairs might elicit increased firing rates during pitch discrimination in some neurons and in localization in other neurons, resulting in no net effect. Under this hypothesis, we would expect that the absolute value of the difference between blocks, across neurons, would be significantly greater for that stimulus pair versus the other stimulus pairs. However, a one-way Kruskal-Wallis test across stimulus pairs again revealed no significant difference: all stimulus pairs yield equivalent absolute differences in evoked firing rate across neurons.
FIGURE S6
Figure S7. More information on electrical disruption of mPFC, related to Figure 7

A) Effect of electrical disruption of mPFC on performance for all individual sessions from three rats (red: Z1, yellow: Z2, green: Z3). Each line connects the performance on control trials (left) and disruption trials (right) within the same session. Plus marks indicate a significant difference, which was an impairment in the mean performance in every case (p < 0.05, Fisher’s exact test). Trials are grouped according to go and nogo in both blocks. Asterisks indicate trial types for which the effect of disruption was significant across sessions (p < 0.05, binomial test on the number of sessions showing impairment).

B) Impairment caused by electrical disruption by stimulus pair, averaged across sessions for each rat. Error bars: SEM across sessions. Impairment is defined as the difference between performance on control trials and disruption trials. Colors represent individual rats, following (A). Two rats (Z1 and Z2) showed a significant decrease for RIGHT+HIGH during pitch discrimination across sessions (p < 0.05, binomial test on the number of sessions showing impairment). One rat (Z3) showed a deficit for both nogo sounds during localization.
Figure S8. Behavior of the model over parameter space, related to Figure 8
A) The model described in Figure 8 operates well over a wide range of parameters. We quantify its performance as the strength of the task signal (x-axis in Figure 8B) necessary for the model to reach a criterion performance of 80%. As in Figure 8B, we normalize this task signal strength to the sensory noise (the variance of the Gaussian noise that was added to each A1 neuron’s activation during the simulation). This plot shows a color-coded map of the strength of the task signal necessary to reach criterion, for various values of N (x-axis) and of the sensory signal-to-noise ratio (y-axis), defined as the strength of the tuning of each neuron to the variance of the same Gaussian noise mentioned above. Darker colors are good: they indicate that a weak task signal was sufficient to reach criterion; lighter colors indicate that a stronger signal was necessary. It is desirable for the model to operate with as weak a task signal as possible, because when the task signal becomes quite strong it overwhelms the sensory input and the performance drops again. We found that larger networks invariably performed better, as expected. For a given network size (any column of the figure), the performance was stable across a certain range of SNRs, but quickly degraded below a certain SNR cutoff. In the white region (small network sizes and/or very poor sensory signal-to-noise ratio), the model never reaches criterion performance.

B) Similar to panel A, but here we show the maximum magnitude of the task signal before performance falls below criterion again. Here, lighter colors (as in the upper right corner) are good, because they indicate that the network performs well over a wide range of task signal strengths; this is associated with large network sizes and/or high sensory signal-to-noise ratios.

FIGURE S8
Supplementary Experimental Procedures

Behavioral parameters
Each stimulus was 250 ms in duration. The warbles were frequency-modulated tones, centered at 6 KHz (LOW) and 16 KHz (HIGH) with a 10 Hz modulation frequency of amplitude 0.07 octaves, and presented with equal intensities (65 dB SPL) from both speakers. The white noise bursts were of approximately equal power at all frequencies between 5KHz and 50KHz, decaying rapidly outside of this range; the total acoustic power over this range was 55 dB SPL, delivered from only one speaker at a time (LEFT or RIGHT). We used a Lynx L22 sound card to convert digital signals to analog voltages, and Fostex FT17H tweeters to produce the sound.

In a subset of sessions, we presented task-irrelevant natural sounds during epochs when the rat was not initiating trials for the purpose of probing receptive fields. These probe sounds were terminated as soon as the rat began performing the task again. We did not observe any correlation between the neural results presented in this study and the receptive fields estimated from responses to these natural sounds.

Surgical implantation
Rats were anesthetized using ketamine/xylazine and isoflurane as necessary to maintain a deep anesthesia, assessed using toe pinch. Skin and fascia were resected from the midline and the skull cleaned. Titanium screws (Small Parts) were inserted into each cranial plate. Two additional stainless steel screws, inserted into the left and right sides of the occipital plate above the cerebellum, served as separate reference and ground signals for the recordings. N-3 flux (LA-CO Industries Inc.) was used to solder wires onto these screws before implantation; after implantation, these wires were soldered to the reference and ground inputs on the microdrive.

Craniotomies were performed directly dorsal to the target areas (A1: 5.25 mm posterior and 6.5 mm left from bregma; PL: 3.0 mm anterior and 1.0 mm left from bregma). The dura was removed and the tetrodes gradually inserted into each region. The craniotomy was filled with agar, which surrounded and protected the tetrodes. Methyl methacrylate (Teet’s, Henry Schein) was used to affix the entire drive to the skull and screws.

Finally, the tissue was flushed thoroughly with sterile saline and sutures were used if necessary to seal the skin around the implant. Aseptic technique was maintained throughout the surgical procedure and the tetrodes themselves were disinfected before implantation. Post-operatively, the rat was given buprenorphine and/or meloxicam to provide analgesia and its health and weight were monitored twice daily. Once the rat had fully recovered, the behavioral task was resumed, concurrent with electrophysiological recording.

Analysis of possible confounds: waveform variation and firing rate drift
In addition to the standard spike-sorting procedure for identifying stable units, we also performed an extra analysis to check the quality of our data and to ensure that the hold period effect could not be due to sorting errors arising from small variations in spike waveform shape between blocks. For each neuron, we identified the sub-cluster of sorted spikes that occurred during the localization hold period, and calculated the Mahalanobis distance (in the first four PCA feature dimensions) between this subcluster and the full cluster consisting of all spikes from the same unit. We assessed the significance of this distance by randomly permuting the labels on the subcluster and the full cluster 2000 times, and calculating the probability of observing a distance less than or equal to the true distance. We repeated this analysis for the other block (pitch discrimination). We discarded the neuron from analysis entirely if the subcluster in either block was significantly more separated from the full cluster than the permuted subclusters were (p < 0.05, permutation test). We also repeated the analysis with simulated Gaussian sub-clusters of the same size as the actual sub-clusters and derived the distribution of mean-squared distance from the cluster center, again rejecting any neuron whose sub-cluster exceeded the 95th percentile of this distribution in either block.
Additionally, we also considered the possibility that a spurious hold period effect could arise from a slow increase or decrease in firing rate over the entire session, perhaps due to drift or motivation, even though the multiple switches between blocks within each session made such a possibility unlikely. We reasoned that, if this were true, then when taking block number into account the difference between block types should no longer be significant. We fit a linear model to the square root of the spike count in the hold period on each trial, using both block type (localization or pitch discrimination) and block number (1, 2, 3, …) as predictors. (Square root is a variance-stabilizing transform for Poisson counts.) We assessed the significance of each predictor with ANOVA. Any neuron that showed a hold period effect according to the analysis described in the text, but that failed to show a significant effect of block type or failed the overall F-test (p > 0.05) was discarded from the analysis. 8/231 neurons (combined across brain regions) were discarded for this reason.

For this ANOVA we used the type-III sum of squares. For all ANOVA analyses in this study, we avoided using type-I sum of squares because we found it to be much more sensitive to unequal trial counts (e.g., more hits in one block than in the other).

**Power analysis**

We analyzed the statistical power of our methods (unpaired Mann-Whitney U-test on the spike counts across blocks) on simulated Poisson counts. We determined the total spike count had to be at least 20 spikes to detect a change between blocks; therefore neurons with fewer total spikes than this in the hold period of all trials combined were discarded from our analysis. For the typical trial counts in our dataset, we would not be able to detect any increase that was less than a doubling of the firing rate for a neuron with this minimum firing rate (though the method becomes much more sensitive at higher firing rates). For this reason, selection rule encoding could be even more common than we have shown. Also, due to the fact that some of our A1 data was collected in animals for which the hold periods were shorter and the trial counts lower, we have less statistical power in that portion of the dataset.

**Calculation of evoked responses**

For each neuron, the spike times on each trial were smoothed with a Gaussian kernel with 1 ms standard deviation. For every 0.5 ms time bin after stimulus onset, the distribution of smoothed spike counts was compared to the combined distribution of all 0.5ms time bins in the 50 ms preceding stimulus onset with a Mann-Whitney U test. The first window of contiguous time bins that were all significantly greater than the spontaneous rates was defined as the onset response window. Windows of less than 1 ms were discarded because these neurons emitted far too few spikes to analyze statistically. A few neurons showing atypical auditory onset responses (e.g., 4/108 showed a slow build rather than a short-latency peak) were discarded because we were concerned that their activity might be driven by the decision rather than the stimulus.

For the average evoked response plotted on the y-axis of Figures 6C and 6D, we used a bootstrap procedure to draw equally from each stimulus, separately for each block. This procedure accounts for differences in the proportion of each stimulus type across blocks, arising from random chance or from better performance on some stimuli than others (e.g., better performance on go than on nogo, Figure S2A) since only correct trials were included for this analysis.

**Changes in evoked response not explained by changes in baseline**

We asked whether there were any additional changes in evoked response, above and beyond what could be explained by pre-stimulus effects, by subtracting the block-specific baseline firing rate from the evoked response on each trial and then repeating the bootstrap procedure described immediately above. We found that a small population (6/43, or 14%) of neurons increased their evoked response significantly (p < 0.05 from the overlap of the bootstrapped distributions in each block), above and beyond any baseline changes. (Another analysis in which we directly compared across blocks the number of spikes emitted in response to each
stimulus individually yielded similar results, as did a stimulus*block ANOVA on each neuron.) However, unlike the populations contributing to the other results in the paper, these neurons were predominantly (4/6) observed in a single animal (Rat 1). This was the rat that had the most difficulty with pitch discrimination, and in these neurons the firing rate was higher during pitch discrimination. One possibility is that the greater difficulty this rat had with one block led to this block-specific increased in evoked rate; our other rats were more evenly matched in performance between blocks.

Disruption of mPFC by electrical stimulation: parameters used
Differences in preparation (i.e., tip geometry, electrode material, stimulus delivery equipment, brain region) and uncertainty about the exact biophysical mechanisms underlying electrical microstimulation make it difficult to estimate the volume of disrupted tissue with certainty. In a previous study, simultaneous calcium imaging of 200us 25uA pulses in rodent V1 (Histed et al., 2009) primarily affected tissue within 50 microns; however, effects were visible throughout the imaging plane (>250um). However, there is no consensus on whether the important parameter for determining the stimulated volume is current (Histed et al., 2009), total charge transfer (Butovas et al., 2003), or charge density per unit area of electrode (Logothetis et al., 2010). Nonetheless, it remains clear that microstimulation affects primarily tissue near the electrode tip and secondarily tissue at larger (millimeter-scale) distances, perhaps by activating axons.

Our goal was to disrupt mPFC, not necessarily to silence or activate it. In fact, microstimulation produces short-latency activation, followed by a long-lasting (hundreds of milliseconds) suppression of neural activity (Butovas et al., 2003). The relative magnitude of the activation and suppression depends in large part on the stimulation frequency. Stimulation in the 5Hz – 40Hz range produces synchronized pulses of neuronal activation on a background of strong suppression; stimulation at 10 Hz in particular produces a complex interaction of activation and suppression (Butovas et al., 2003). Moreover, experiments in the primate visual system (Logothetis et al., 2010) using much higher currents but a larger surface area electrode (250uA, 200us, 6000 square microns) determined 10 Hz to be an optimal frequency with which to stimulate the visual thalamus in order to maximally suppress the corresponding retinotopic location in V1 as well as its downstream projection targets. Therefore, in order to maximize the disruption induced by our stimulation, we chose a 10 Hz stimulation frequency in our experiments.

Note also that nearby neurons in mPFC are almost certainly less homogeneous in their functions and coding properties than nearby neurons in many visual sensory areas, such as MT. Thus, even if we used a stimulation protocol that strongly activated a large number of nearby neurons in mPFC, it would most likely still result in disruption of mPFC function, rather than produce the sort of coherent signal that has been used to bias decisions in primates (Salzman et al. 1992), for example.

We estimate the exposed surface area of each MicroProbes electrode (used in Z2 and Z3) as a cone of radius 4.5 um and height 32.5um, for a total of 463 square microns. Based on impedances, we expect the exposed area of each FHC electrode (used in Z1) to be comparable.

Disruption of mPFC by electrical stimulation: behavioral results
We sometimes used the same stimulation protocol during epochs in between trials when the rats were not behaviorally engaged in the task. We never noticed any overt behavioral response to stimulation under these conditions. Besides the impairment described in the main text, the only additional effect we observed during behavior was that rats appeared to have more trouble completing the center poke. The sound does not play until the rat holds for a random duration between 250ms and 350ms. Shorter ("failed") center-pokes do not initiate a trial. Consistent with the proposed role of the mPFC in behavioral inhibition (impulse control) and in estimating temporal duration, we noticed that rats exhibited more failed center-pokes during disruption (data not shown), especially at higher current levels. Typically they did not go to the choice port after a failed center-
poke (that is, the stimulation did not directly elicit a choice motion); they simply repeated the center-poke until successfully initiating a trial.

For some sessions, the disruption caused a significant increase in the number of “wrong-port” responses — trials in which the rat went to the choice port associated with the other block (data not shown). This suggested a possible specific deficit in stimulus selection, or in the memory of the rules for the current block, but the effects were insufficiently consistent to draw firm conclusions. We only included sessions for which the performance on control trials was significantly above chance, using the same definitions of chance performance as we did previously for the non-stimulated animals (see: “Chance performance on the task” in the Methods section of the main text).

We observed that rats appeared to be particularly impaired on the “congruent” nogo stimulus RIGHT+HIGH in one or both blocks (Figure S7B). This is interesting because this stimulus should, in theory, be the least ambiguous stimulus of all: it always requires a nogo response in either block. For this reason, there was no significant increase in the proportion of trials on which the rat gave the response that would have been appropriate in the other block, as one might have expected were stimulus selection the only cognitive ability that was affected. Future experiments will be needed to disentangle the role of the mPFC in holding the center port, interpreting the stimulus, and producing the correct motor act.

**Simulated network model**

In our model, simultaneously presented pairs of stimuli produce activity patterns in a population of randomly tuned A1 neurons. The A1 neurons are divided arbitrarily into two subpopulations. The first subpopulation is used for task 1 (i.e., localization) and projects to two downstream command neurons representing the possible responses in that block (“go left” and “nogo”). Similarly, the second subpopulation projects to two different command neurons representing the possible response in the other block (“go right” and “nogo”). This is described by the following equations:

\[
S \cdot \text{STRF}_1 \cdot \text{H}_1 = \text{B}_1 \\
S \cdot \text{STRF}_2 \cdot \text{H}_2 = \text{B}_2.
\]

S: “Stimulus”. Binary matrix with shape \((N_{\text{trials}}, 4)\) encoding the presence of two out of four possible sounds (left, right, low, or high) for each stimulus pair in \(N_{\text{trials}}\) test trials.

STRF\(_1\) and STRF\(_2\): “Spectrotemporal receptive field” or “receptive field” matrices with shape \((4, N_{\text{neurons}})\) encoding the fixed tuning for each of the four sounds of each neuron in each subpopulation. These are set randomly at the beginning of the simulation and subsequently held constant.

H\(_1\) and H\(_2\): “Projection” matrices with shape \((N_{\text{neurons}}, 2)\) representing the projection from each subpopulation to the two command neurons used by that subpopulation. During training, optimal values are chosen for these matrices.

B\(_1\) and B\(_2\): “Output”. Binary matrices with shape \((N_{\text{trials}}, 2)\) representing the correct command neuron activation (i.e., correct motor response) for each stimulus pair in the test set.

During training, we choose random stimulus pairs for the test trials and we fix the values of B\(_1\) and B\(_2\) to be the correct output activations for tasks 1 and 2, respectively. We next solve for the projection matrices H\(_1\) and H\(_2\) that produce the minimum squared residual, using a non-negative least squares algorithm (scipy.optimize.nnls) that produces only positive (excitatory) weights. Note that each subpopulation is essentially trained separately — the first subpopulation produces the correct responses (B\(_1\)) for task 1; the second subpopulation produces the correct response (B\(_2\)) for task 2.
During the test phase, we generate a new set of test stimulus pairs but we keep STRF$_1$ and STRF$_2$ fixed at the values chosen during training. The entire network’s choice on each test trial is defined as the command neuron with maximal activation (“winner-take-all”). The performance of the model is assessed as the percentage of the time that the network’s choice matches the correct response for that test stimulus.

Critically, during the test phase we add a “task signal” to one of the subpopulations to bias the entire network towards the task encoded by that subpopulation. Specifically, during task 1, we calculate the command neurons’ activations as follows:

\[
(S \ast \text{STRF}_1 + T) \ast \text{H}_1 = B_1 \\
(S \ast \text{STRF}_2) \ast \text{H}_2 = B_2.
\]

Here, T is a matrix with dimensions \((N_{\text{trials}}, N_{\text{neurons}})\) in which every element is the same, positive value. This corresponds to adding a small positive amount on every trial to the activation of every neuron in subpopulation 1. Because H$_1$ is strictly positive, this translates into a positive increase in the activation of both command neurons in B$_1$. (During task 2, we perform the opposite manipulation by adding T to the second subpopulation. For simplicity, this is represented as a negative task signal in Figure 8 of the main text.)

For a given SNR, as the task signal increases in strength, the model’s performance eventually drops to 50%. This is because the size of the task signal begins to dominate the sensory input (the “overdriven” regime). In this regime, the model still produces responses that are appropriate to the block (i.e., it does not “go to the wrong port”) but the responses are no longer related to the sensory input. In terms of the equations above, this occurs because both command neurons encoded by B$_1$ are increasingly activated in task 1; however, only one of these two command neurons actually encodes the correct response. When the task signal is sufficiently strong, the task-specific activation of the neurons in B$_1$ swamps the stimulus-specific activation of the correct neuron, and the model chooses at random between the two neurons in B$_1$ (the two possible responses during that block.)

When the task signal corresponding to task 1 is activated, but the model is assessed on its performance on task 2, it performs poorly (as expected). These data are shown in Figure 8B of the main text: negative values of the task signal correspond to activation of the “wrong” network. This corresponds to doing localization instead of pitch discrimination in the real task; in such a case, only 25% of trials will be correct: those which present the RIGHT+HIGH stimulus which always means nogo.

For a network size \(N = 640\), and at very low SNRs <1%, our model reached the overdriven regime before it ever produced good performance. We measured the minimum and maximum values of the task switch signal that produced good performance (>80%) over a range of network sizes and SNRs. Increasing the size of the network increases the working range of the model by decreasing the minimum task signal that is necessary to reach criterion performance and increasing the maximum acceptable task signal before reaching the overdriven regime (Figure S8).

**Rule-encoding is not dependent on the motion history of the animal**

We asked whether the effects could be explained by the motion history of the animal preceding center-poke entry. We reasoned that the rat’s movement history before each trial would strongly correlate with the choice on the previous trial. In particular, if the rat had just completed a successful nogo trial, it was likely to have remained relatively motionless in the center port; on trials following a successful go trial, it was likely to have just moved from the choice port to the center port. When we analyzed only trials following a successful nogo and found that the rule-encoding remained the same: the firing rate was still elevated during the preferred block and suppressed during the non-preferred block (Figure S5B). This demonstrates that the rule encoding cannot be purely an effect of the motion to the center port.
**Video analysis of preparatory head positioning — methods**

We recorded video of all behavioral sessions using an infrared camera located directly above the rat. We hypothesized that the rat might use a different posture in each block since the two reward ports were in different locations. We took the video frame closest to the center of the hold period on each trial and manually scored the position of each ear. We did this by asking a human observer (CR), who was blind to the block and outcome of each trial, to click on each ear in the frame from every trial and record the position of the clicks.

Using this time consuming scoring procedure, we analyzed the three sessions during which we recorded the most rule-encoding neurons (one session each from Rats 2, 4, and 6). Using the position of each ear, and knowing that the nose was located in the center port during this interval, we were able to construct the center position and azimuthal angle of the head relative to the behavior box. We found a prominent correlation between the head angle and the block in all analyzed sessions (e.g., Figure S3I, S4I). The mean difference between blocks was 26.9, 36.6, and 24.1 degrees in each of the three analyzed sessions.

Because head angle is correlated with block, and because we analyzed rule-encoding neurons for which, by definition, the firing rate correlated with block, it stands to reason that head angle correlates with firing rate. In an example neuron (Figure S3J, S4J), the firing rate is highly significantly correlated with block (black trend line). However, for this neuron the firing rate is not correlated with head angle within each block separately (red and blue trend lines), implying that this is a mere side effect of the fact that the neuron is encoding block.

To estimate the relative contributions of head angle and block identity in driving the response of each neuron, we used multiple-regression/analysis-of-variance (ANOVA), a standard technique for estimating the contribution of each, possibly correlated, variable to an observed response (Sokal and Rohlf, 2003). This technique constructs the optimal linear model for predicting neural firing rate using both head angle and block.

The amount of variance that each factor explains in the optimal linear model is taken as an estimate of the amount that it contributes to the actual response. After defining our model, we discuss further the reasoning behind this argument, assuming linearity throughout; we also present simulations we conducted to validate that the approach is accurate and effective on datasets like this one.

Our model was defined as follows:

\[
\text{Response} = B_0 \times \text{Block} + B_1 \times \text{HeadAngle} + \text{residual}
\]

Here, Block is a binary variable representing pitch discrimination or localization, HeadAngle is the measured head angle on each trial, and Response is the square root of the spike count on each trial. (Square root is a variance-stabilizing transformation for Poisson-like spike counts: it ensures the variance is homogeneous and independent of mean firing rate.)

Multiple regression yields the optimal values of \(B_0\) and \(B_1\), in the sense of minimizing the squared sum of the residuals. These coefficients may be seen by inspection to give the contribution of one variable while the other variable is held constant, which is exactly what we want to estimate. Analysis-of-variance (ANOVA) yields the explained variance (EV) of each variable. In this case it is simply equivalent to:

\[
\text{EV[Block]} = \text{var}(B_0 \times \text{Block}) \\
\text{EV[HeadAngle]} = \text{var}(B_1 \times \text{HeadAngle})
\]

Finally, the FEV is the explained variance of each individual variable, divided by the sum of explained variances over all variables.

How does multiple-regression/ANOVA assign EV when the factors are themselves correlated? Technically, this is done by inverting the covariance matrix in the so-called normal equations:

\[
B = \text{inv}(X^T X) X^T Y
\]
where X is the matrix of factors, Y is the response, and B is the matrix of coefficients $B_0$ and $B_1$. That is, shared variability between the factors is removed (a process called “whitening”) before calculating the coefficient of each factor. Thus the FEV is a measure of the explanatory power uniquely available to each factor in the model, after removing all shared explanatory power from the factors (Sokal and Rohlf, 2003).

An alternative calculation yields the same coefficients but provides a different intuition. To calculate the coefficient ($B_i$) for each variable, we first hold out that variable and find the best fit between the other variable and the response. We next fit the residuals (unexplained part of the response) to the held-out variable. The coefficient $B_i$ thus obtained is the same $B_i$ that is yielded by multiple regression. We then repeat for all other variables. In this way it may be seen that the coefficient of each variable (e.g., HeadAngle) is determined only by how well it predicts the part of the data that cannot be linearly predicted by Block, and vice versa.

Hence, the FEV is not only a measure of the variable’s contribution to the optimal linear model, but also a measure of that variable’s unique explanatory power divided by the total unique explanatory power of all variables, when considering the space of all possible linear models with those variables.

Reference:

Video analysis of preparatory head positioning — results
Across our population of rule-encoding neurons, far more of the variability was explained by block than by head position (Figure S3K, L; S4K, L). The fact that some neurons do encode head angle, not block, according to this analysis is consistent with previous results from McNaughton and colleagues. It is also a proof of principle that, in at least some cases, the head angle scoring procedure is sufficiently sensitive to uncover these effects. Nonetheless, the major conclusion of this analysis was that, for most neurons in both brain regions, head angle plays only a small role in the effects we observed.

[Technical note: we also considered a 2-way model that included an interaction term, which allowed for the possibility of different head-angle encoding in each block. The general conclusion however remained the same: block explained most of the variability in most neurons. Finally, one A1 neuron was excluded from this analysis because the least-squares fit was too poor (p-value for the F statistic was greater than 0.05); however, this neuron had very little sensitivity to head angle and including it would not have changed the conclusions. We were still able to consider this neuron using the analysis in the section “Posture equalization”, which does not rely on a least-squares model.]

We chose to focus on head angle because this was the most prominent and easily quantifiable postural difference between blocks and is therefore likely to be correlated with preparatory motor activity in general. No analysis can rule out the possibility that PFC is actually encoding some unknown, subtle difference in behavior between blocks. However, the effects we observe favor the hypothesis that activity of PFC neurons primarily encodes the current task. It is possible that a side effect of this difference in cognitive state is a block-specific difference in motor planning and execution; these motor effects would thus correlate with PFC activity even though they may not be directly encoded by PFC activity. In fact, even our neurons that seem to be encoding primarily head angle may actually be encoding cognitive state: trials on which the rat most strongly plans to perform localization might also produce the strongest preparatory actions.

Verification of ANOVA on simulated posture-encoding neurons
We asked whether our multiple-regression / ANOVA approach was capable of detecting neurons that encoded head angle and not block. Although we did detect a minority of neurons with this property, suggesting that our
method is indeed capable of this, we further investigated this with simulation. For each rule-encoding neuron for which we had video tracking data, we designed a model neuron with a firing rate given by:

\[
\text{Response} = M \times \text{HeadAngle} + \text{Noise}
\]

Here again, as with all of our regression analyses, we define the response as the square root of the spike count in order to stabilize the variance. This ensures homogeneity of variance independent of mean firing rate.

To ensure that the simulation best matched the real data, we defined \( M \) as the slope of the best linear fit between the response and the head angle. We then ran 1000 simulated experiments, each with the same head angles as in the real data but with the noise randomly drawn from a Gaussian distribution. We defined this noise distribution such that the simulated response of the neuron had the same mean and variance as the response of the actual neuron.

We used the same multiple-regression/ANOVA analysis on these model neurons as we did on the real neurons, including both Block and HeadAngle as factors in the model. This analysis correctly identified all model neurons as primarily driven by head angle (Figures S3Q, S4Q): the FEV of head angle was always greater than 50% in the model neurons. The median FEV of head angle was 87.2% in A1 and 86.7% in PFC, compared with 17.0% and 5.9% in these brain regions in the real data.

Note that Block still correlated with HeadAngle just as before, meaning the model could have used either factor to predict the response. However, the analysis correctly assigns high FEV to HeadAngle here, in striking contrast to the high FEVs assigned to Block in the real data. We conclude that this approach is capable of detecting neurons that primarily encode head angle.

**Verification of ANOVA with correlated regressors**

Identifying the relative contributions of two correlated variables to a third variable is an inherently difficult problem because the predictive power of the shared variability is not uniquely due to either factor. We verified by simulation that our original procedure (multiple-regression/ANOVA) could accurately determine the fraction of explainable variance (FEV) under a variety of conditions: correlated regressors, binomial factors such as block identity, and discrete output similar to spike counts.

We took the measured head angle and block identity from the video recordings of the sessions and constructed a model neuron with a response given by:

\[
\text{Response} = B_0 \times (\text{HeadAngle}) + B_1 \times (\text{BlockIdentity}) + \text{Noise}
\]

\[
B_0^2 + B_1^2 = 1
\]

HeadAngle and BlockIdentity were first normalized to zero mean and unit variance. We used Gaussian noise with zero mean and variance of 0.5, though the results were insensitive to the exact noise level provided there was some noise present. Thus, by construction we know the true FEV of HeadAngle and BlockIdentity to be \( B_0^2 \) and \( B_1^2 \), respectively. We used our previous least-squares/ANOVA analysis to derive estimates of the FEV, and compared these estimates to the known, true value. We found that the ANOVA analysis estimated the FEV of both factors with near zero bias (Figures S3R) over the full range of values of \( B_0 \) and \( B_1 \) between 0 and 1. That is, it properly accounted for the relative contributions of head angle and block identity. The error bars on the estimates were determined by the noise level in the simulation.

We extended this analysis by replacing the empirical distributions of the factors with simulated sessions, in which HeadAngle was drawn randomly from a normal distribution and BlockIdentity was a binary variable constructed to have a certain correlation with HeadAngle (Figure S3S). (Technically this was done by first assigning BlockIdentity by thresholding HeadAngle, which produces maximal correlation, and then flipping
BlockIdentity randomly until the desired correlation was reached.) In these simulated cases we also found a
good match between the true and estimated values of FEV (example $r=0.72$ Figure S3S; similar results were
found for all values of $r$ over its entire range).

Finally, we obtained very similar results when we converted the output of the model neuron to a rectified,
discrete quantity that more closely resembled spike counts (Figure S3T). Specifically, we first exponentiated
the output, $x$, to obtain a positive quantity analogous to the expected spike count: $y = 2^{x/x_0}$ for some reference
value $x_0$. We then obtained simulated spike counts by drawing random integers from a Poisson process with
mean count set to $y$. In this case we found it necessary to use a square-root variance-stabilizing transformation
on the spike counts, just as we previously did with the real data. Throughout, we were able to achieve good
results using either type-2 or type-3 (but not type-1) sum-of-squares in ANOVA.

We conclude that ANOVA is a robust tool for estimating the fraction of explainable variance under a wide
variety of conditions: correlated regressors, binary regressors, and Poisson-like quantized output. We found
that our ANOVA analysis performed well for all of these types of data, including when using simulated spike
counts and real values of head angle and block identity.

Posture equalization by trial selection
We also conducted an additional analysis that directly controls for the effect of head angle by trial selection. In
brief, we considered a subset of correct trials from each block (“posture-equalized trials”), selected such that
the mean head angle was the same in both blocks, and compared the rule encoding on those trials with the
rule encoding on the rest of the trials. If posture accounted for the entire difference between blocks, then the
difference between blocks should disappear on the posture-equalized trials. For most neurons in either brain
region, we found the rule encoding to be similar on posture-equalized trials, suggesting that head angle plays
only a minor role overall.

We present this analysis in three parts. First, we show that the distribution of rule encoding across neurons on
posture-equalized trials is significantly different from what would be expected by random chance, and is
broadly similar to the result from the rest of the trials (top and middle panels of Figures S3M, S4M). Second,
we show that the mean magnitude of rule encoding is not significantly altered on the posture-equalized trials
(Figures S3N, S4N). Third, we show that individual rule encoding neurons rarely change their rule encoding on
posture-equalized trials (Figures S3O, S4O). All results hold for both mPFC (Figure S3) and A1 (Figure S4).

We first show the distribution of rule encoding, quantified as the logarithm of the ratio of the firing rates in pitch
discrimination over localization, separately for posture-equalized and all other trials (Figures S3M, S4M).
These distributions appear broadly similar, suggesting that rule encoding is not strongly altered by posture
equalization. If rule encoding were entirely due to postural differences, then the distribution should be tightly
clustered around zero on the posture-equalized trials, but we do not observe this. To demonstrate this, we
randomly shuffled the block labels on the posture-equalized trials 10000 times and calculated the rule
encodings on each shuffle. The bottom panels of Figures S3M and S4M show the distribution of the
(necessarily spurious) rule encodings thus obtained: it is narrowly clustered around zero, as expected, and
appears much narrower than the top distribution of posture-equalized rule encoding.

The spread of the distribution of rule encoding across neurons is a measure of the “population rule encoding”,
which we will refer to as PRE; a broad distribution indicates the presence of many neurons that strongly prefer
one or the other block. We quantified the PRE as the median absolute deviation (MAD) of the distribution of
rule encoding. The MAD is a statistical estimator of the standard deviation that uses the median and is
therefore more robust to outliers (Quian-Quiroga et al., 2004). The PRE was 0.471 in PFC and 0.318 in A1 on
the posture-equalized trials; the fraction of shuffles that yielded a greater PRE was <0.005 in both brain
regions. Thus, the rule encoding was significantly greater on the posture-equalized trials than it was on the
randomized shuffles, and we may reject the hypothesis that the observed rule encoding on posture-equalized
trials is simply due to random variation. We note that this conclusion was the same when we used the standard deviation instead of the MAD (data not shown).

(Technical aside: on a small fraction of the shuffles, we obtained zero spikes in one or the other block, which yielded a rule encoding of infinity. This occurred with probability <0.005. Such points cannot be shown on the distribution in Figures S3M and S4M. One advantage of the MAD over the standard deviation is that it yielded a robust estimate of PRE across shuffles, even in this rare and extreme circumstance.)

Second, we found that the mean magnitude of rule-encoding was not significantly altered by postural equalization (Figure S3N, S4N). We compared the absolute value of rule encoding for each neuron before and after posture-equalization using a paired Mann-Whitney test and found no significant difference.

Finally, we plotted the rule encoding for each individual cell before and after posture equalization (Figure S3O, S4O) and asked how often the preferred block changed after posture equalization. (Note that “preferred block” here simply means the block with greater firing rate, while in the rest of the text it is the block with significantly greater firing rate. We did not use statistical thresholding here because the sample size is smaller after posture equalization, decreasing statistical power.) If rule encoding were entirely due to posture differences between blocks, then posture equalization should remove any difference between block; thus the preferred block after posture equalization would be random and, 50% of the time, different than the preferred block in the rest of the trials. In fact we observed that only 1/16 PFC cells and 0/8 A1 cells switch their preferred block after posture equalization. The p-value (two-tailed binomial test) of this result is 0.008 for A1 and 0.0005 in PFC. Thus, we may reject the null hypothesis that posture equalization removes any difference between the blocks.

In summary, posture equalization appeared to non-significantly weaken the rule encoding, but it did not remove it. This weakening effect is to be expected based on our result that head angle explained some of the firing rate, primarily in a small minority of our neurons (Figures S3K, S4K), and also based on previous reports that some neurons in mPFC encode head angle (Cowen and McNaughton, 2007). However, we emphasize that even after posture equalization, the rule encoding was largely preserved, in terms of its population distribution (Figures S3M, S4M), its mean magnitude (Figures S3N, S4N), and its sign in individual neurons (Figures S3O, S4O).

Acknowledgements
The authors would like to thank Mohammad Dastjerdi and Christian Fernandez for developing the initial rodent behaviors in the lab and managing behavioral training; Alfonso Apicella for helping set up the lab; Tony Zador, Uri Livneh, Erin Rich, Nicholas Steinmetz, Anita Devineni, and Aaron Koralek for comments on the manuscript; Sarah Kochik, and Vuong Vu for managing behavioral training; Trevor Dolinajec, Lily Lin, Danielle Jobe, and Yassi Sabahi for electrode fabrication; Yuliy Tsank, Sharon Lin, Ambika Rustagi, Amish Shah, Carol Pham, Francine Foo, Helene Moorman, Jason Murphy, Jason Zhang, Josh Chong, Kelly Clancy, Tanling Hsu, Pooja Upadhyaya, and Tommy Li for help with rat behavior; and Vivek Ayer for technical assistance.