JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Systems/Circuits

Selective modulation of orbitofrontal network activity during negative occasion setting

Justin L. Shobe[1], Konstantin I. Bakhurin[2], Leslie D. Claar[3] and Sotiris C. Masmanidis[1,2,4,5]

[1]*Department of Neurobiology, University of California, Los Angeles, California 90095, USA.*
[2]*Neuroscience Interdepartmental Program, University of California, Los Angeles, California 90095, USA.*
[3]*Department of Bioengineering, University of California, Los Angeles, California 90095, USA.*
[4]*Integrative Center for Learning and Memory, University of California, Los Angeles, California 90095, USA.*
[5]*California Nanosystems Institute, University of California, Los Angeles, California 90095, USA.*

DOI: 10.1523/JNEUROSCI.0572-17.2017

Received: 1 March 2017

Revised: 18 July 2017

Accepted: 24 July 2017

Published: 28 August 2017

Author contributions: J.S. designed research; J.S. and L.D.C. performed research; J.S., K.I.B., and S.C.M. analyzed data; J.S., K.I.B., L.D.C., and S.C.M. wrote the paper.

Conflict of Interest: The authors declare no competing financial interests.

We thank D.V. Buonomano and V. Goudar for assistance with decoding analysis. K.I.B. acknowledges support from the Ruth Kirschstein National Research Service Award NIH T32-NS058280. S.C.M. acknowledges support from a 2014 McKnight Technical Innovations in Neuroscience Award, NIH DA034178, and NIH NS100050.

Corresponding authors: Justin L. Shobe (jshobe@gmail.com), Sotiris C. Masmanidis (smasmanidis@ucla.edu), University of California, Los Angeles, Los Angeles, CA 90095

Cite as: J. Neurosci ; 10.1523/JNEUROSCI.0572-17.2017

Selective modulation of orbitofrontal network activity during negative occasion setting

Abbreviated title (50 character maximum)

OFC encoding of negative occasion setting

Authors

Justin L. Shobe[1], Konstantin I. Bakhurin[2], Leslie D. Claar[3], and Sotiris C. Masmanidis[1,2,4,5]

Affiliations

[1]Department of Neurobiology, University of California, Los Angeles, California 90095, USA.

[2]Neuroscience Interdepartmental Program, University of California, Los Angeles, California 90095, USA.

[3]Department of Bioengineering, University of California, Los Angeles, California 90095, USA.

[4]Integrative Center for Learning and Memory, University of California, Los Angeles, California 90095, USA.

[5]California Nanosystems Institute, University of California, Los Angeles, California 90095, USA.

Corresponding authors

Justin L. Shobe (jshobe@gmail.com)

Sotiris C. Masmanidis (smasmanidis@ucla.edu)

University of California, Los Angeles

Los Angeles, CA 90095

Number of pages: 31

Number of figures: 5

Number of words: 8590

Abstract: 250

Introduction: 609

Discussion: 880

Conflict of interest

The authors declare no competing financial interests.

Acknowledgments

We thank D.V. Buonomano and V. Goudar for assistance with decoding analysis. K.I.B. acknowledges support from the Ruth Kirschstein National Research Service Award NIH T32-NS058280. S.C.M. acknowledges support from a 2014 McKnight Technical Innovations in Neuroscience Award, NIH DA034178, and NIH NS100050.

Abstract:

Discrete cues can gain powerful control over behavior in order to help an animal anticipate and cope with upcoming events. This is important in conditions where understanding the relationship between complex stimuli provides a means to resolving situational ambiguity. However, it is unclear how cortical circuits generate and maintain these signals that conditionally regulate behavior. To address this, we established a Pavlovian serial feature negative conditioning paradigm, where male mice are trained on a trial in which a conditioned stimulus (CS) is presented alone and followed by reward, or a feature negative trial in which the CS is preceded by a feature cue indicating there is no reward. Mice learn to respond with anticipatory licking to a solitary CS, but significantly suppress their responding to the same cue during feature negative trials. We show that the feature cue forms a selective association with its paired CS, because the ability of the feature to transfer its suppressive properties to a separately rewarded cue is limited. Next, to examine the underlying neural dynamics, we conduct recordings in the orbitofrontal cortex (OFC). We find that the feature cue significantly and selectively inhibits CS-evoked activity. Finally, we find that the feature triggers a distinct OFC network state during the delay period between the feature and CS, establishing a potential link between the feature and future events. Taken together, our findings suggest that OFC dynamics are modulated by the feature cue and its associated conditioned stimulus in a manner consistent with an occasion setting model.

Significance Statement:

The ability of patterned cues to form an inhibitory relationship with ambiguously rewarded outcomes has been appreciated since early studies on learning and memory. However, it was often assumed that these cues, despite their hierarchical nature, still made direct associative links with neural rewarding events. This model was significantly challenged, largely by the work of Holland and colleagues, who demonstrated that under certain conditions cues can inherit occasion setting properties whereby they modulate the ability of a paired cue to elicit its conditioned response. Here we provide some of the first evidence that the activity of a cortical circuit is selectively modulated by such cues, thereby providing insight into the mechanisms of higher order learning.

Introduction:

Animals routinely learn to anticipate events by extracting information from their environments. However, this can be particularly challenging when individual cues only provide partial predictive information as is often the case in naturalistic scenarios. In these situations, animals will attempt to use disambiguating 'features' in order to accurately predict outcomes (Schmajuk and Holland, 1998). A good example of this type of learning is feature negative conditioning because behavioral success requires an animal to learn the pattern of cues that best predicts reward (Holland, 1984; Lamarre and Holland, 1987; Bueno and Holland, 2008). In the serial version of this task, animals learn that a single conditioned stimulus (CS) predicts a reward but when this same cue is preceded (with a temporal delay) by a separate feature cue, the trial goes unrewarded (Holland, 1985; 1992). Thus, the single cue elicits anticipatory behavior, but animals withhold their responses when the same cue is presented in feature negative trials. Studies have shown that the ability to conditionally discriminate between rewarded and unrewarded trials can occur in a wide range of species, from insects to humans (Pace et al., 1980; Nallan et al., 1981; Pace and McCoy, 1981; Abramson et al., 2013), and under a variety of stimulus conditions (Holland, 1992; 1997). In the mammalian brain there is evidence that these functions are mediated by specific circuits, including the retrosplenial cortex (Robinson et al., 2011), striatum, and orbitofrontal cortex (Meyer and Bucci, 2016). Despite these studies, there is still a relatively poor understanding of the relationship between feature cues and their associated conditioned stimuli that function to bias behavioral decisions.

There are two contrasting models that attempt to account for how neural circuits solve this problem. One model views the animal's ability to discriminate rewarded and unrewarded trials as a basic function of elemental conditioning, where a CS acquires a positive associative relationship to the reward to promote conditioned responding, and the feature acquires a negative relationship to suppress responding (Rescorla, 1969; Rescorla and Wagner, 1972; Rescorla and Holland, 1977). On trials in which both cues are present, the feature cue's

inhibitory influence simply overrides the CS's excitatory influence, due to the feature cue's direct negative association with the reward representation (conditioned inhibition model). In the opposing model, the feature cue functions as a negative occasion setter that does not make a direct association with the reward representation (Lamarre and Holland, 1987; Holland, 1984, 1989; 1995a). Instead it modulates the ability of the CS to retrieve the reward association by acting as a kind of inhibitory gate (Holland, 1989; 1995a).

To gain mechanistic insight into these opposing models at the level of single-neuron spiking activity, we establish a Pavlovian feature negative conditioning paradigm in head-restrained mice, which is compatible with large-scale neural recordings using silicon-based microprobes. In our task, a CS predicts the delivery of reward, but there is no reward when this CS is preceded by a feature cue (Holland, 1995b; 1995a). We find that mice predominantly solve this task by using a strategy consistent with the second model (negative occasion setting), because the feature acquires the ability to specifically inhibit the reward association of its paired CS (Holland, 1984; 2008). Moreover, we find that neural activity within the OFC is consistent with this model because the feature appears to selectively modulate cue-evoked firing in a manner that correlates with behavioral performance. Finally, we also observe an 'activity silent' state (Stokes, 2015) in OFC network dynamics that could function to relay information during the time gap between the feature and CS cue. To our knowledge this is the first demonstration of a modulatory cortical circuit mechanism that specifically supports the occasion setting model.

Materials and Methods:

Animals and surgical procedures

All procedures were approved by the University of California, Los Angeles Chancellor's Animal Research Committee. Singly housed male C57Bl/6J mice (n = 8, 15-22 weeks old at the time of recording, The Jackson Laboratory) were used in the experiments. Animals underwent an initial head bar implantation surgery under isoflurane anesthesia in a stereotaxic apparatus to bilaterally fix, with dental cement, stainless steel head bars on the skull. After training, animals underwent a second surgery under isoflurane anesthesia on the recording day to make a single craniotomy for acute silicon microprobe recordings. An additional craniotomy was made over the posterior cerebellum for placement of an electrical reference wire. All behavioral training and recording sessions were carried out in fully awake head-restrained animals.

Behavioral task

We started food restriction one week after the initial head bar implantation surgery. Mice were fed daily after each training session to maintain ~90% of their baseline weight whereas water remained freely accessible in the home cage. To begin each training session, we mounted animals on the head bar restraint bracket and placed them on a polystyrene treadmill ball (200 mm diameter, Graham Sweet Studios) that freely rotated in a forward/backward direction. Behavioral training consisted of four successive phases: 1) habituation, 2) odor and air puff conditioning, 3) feature negative conditioning, and 4) behavioral testing and electrophysiology. In the first phase, mice were initially habituated to the head restraint system and trained to consume a liquid reward (5 µL, 10% sweetened condensed milk) delivered by actuation of an audible solenoid valve (Neptune Research). Licking was continuously monitored via an infrared lick meter placed in front of the reward delivery tube (Island Motion). During these sessions, animals were given rewards and exposed to a constant stream of pure air through a tube with a hole positioned in front of the nose (50 rewards per session, 13-21 s inter-trial interval (ITI), 1.5

L/min air flow). After mice learned to lick to at least 90% of the delivered rewards for two consecutive days, we began the second training phase. Mice received trials containing one of either two types of olfactory conditioned stimuli (CS1 or CS2, 1 s duration, 17-29 s ITI), or a mild air puff to the vibrissal pad. The air puff was odorless and thus provided a distinct (from the CS1 and CS2) but highly salient form of stimulus, which has been effectively used in head-fixed mouse behavioral paradigms (Guo et al., 2014). Aromatic compounds (isoamyl acetate in CS1, citral in CS2, Sigma-Aldrich) were diluted 1:100 in mineral oil (Sigma-Aldrich). Air (0.15 L/min) was bubbled through this liquid and combined with the 1.5 L/min stream of pure air. An additional air puff tube (which was separate from the odor delivery tubing system to prevent odors being mixed with the air puff) delivered a pulse of pure air to the vibrissal pad (0.5 s at 0.8 L/min) on the side contralateral to the recording hemisphere. This intensity level did not evoke any noticeable startle response such as blinking. CS1 and CS2 were always associated with reward, which was delivered 2.5 s after odor onset. The 1.5 s gap between the offset of the odor and the reward allows cue-evoked behavior and neural activity to be examined in the absence of potentially confounding reward stimulus signals. The air puff was not followed by any explicit outcome. Animals received 30 presentations of each trial type (CS1, CS2, air puff) in pseudorandom order during daily sessions in the second phase of training. The solenoid valves controlling the olfactory cues were sound-isolated and thus inaudible to the animal. Typically, within two days of training, animals began predicting the delivery of reward following CS1 or CS2 cues by exhibiting anticipatory licking during the interval between the cue and reward. After mice demonstrated anticipatory licking on at least 90% of both CS1 and CS2 trials, we began the third phase of training, in which the air puff was now set to serve as the feature cue. On unrewarded trials the air puff was presented starting 2.5 s before CS onset. The third training phase contained an equal proportion (33%) of $CS1^{+}$, $CS1^{-}$, and $CS2^{+}$ trials presented in pseudorandom order (approximately 100 trials per session; Figure 1B, left). The superscript '+' denotes that a CS was not preceded by a feature cue and was followed by

reward, while the superscript '-' denotes that a CS was preceded by a feature cue and was not followed by reward. The minimum reaction time for animals to initiate anticipatory licking was found to be around 0.5 s. Throughout the manuscript we define correct CS^+ trials as those containing anticipatory licking (when licking occurred between 0.5 and 2.5 s following odor onset), correct CS^- trials as those in which animals withheld licking during this time period, and incorrect CS^- trials as those when animals licked during this time period. When mice achieved at least 90% correct CS^+ trials and less than 10% incorrect CS^- trials, we began the last training phase, comprised of a single session which coincided with electrophysiological recordings. Here we introduced transfer trials (TT) in which the CS2 cue was preceded by an air puff feature cue (a novel pairing) and followed by reward (Figure 1B, right). This last phase consisted of 28% $CS1^+$, $CS1^-$, $CS2^+$ trials, and 15% transfer trials. Since the feature had never been previously associated with CS2, we used these transfer trials to determine which of two models (see Introduction) are implemented by the animals. To calculate the behavioral discrimination score, we subtracted the percentage of incorrect $CS1^-$ trials from the percentage of correct $CS1^+$ trials.

Electrophysiological recordings

One recording was performed per animal with a microprobe containing a total of 256 electrodes divided across 4 prongs that were spaced 0.2 mm apart. An array of 64 electrodes on each prong spanned 1 mm along the dorsal-ventral axis. We recorded from the orbitofrontal region of the prefrontal cortex (2.3 to 2.5 mm anterior, 0.5 to 1.5 mm lateral, -2.0 to -3.0 mm ventral, relative to bregma). The silicon prongs were coated with a fluorescent dye (DiD, Thermo Fisher) prior to insertion, to facilitate post hoc histological reconstruction of the recording sites. Procedures for recording with silicon microprobes are described elsewhere (Shobe et al., 2015). After the recordings, animals were overdosed with isoflurane and perfused with 10% formalin solution (Sigma-Aldrich). The brain was extracted and fixed for a minimum of 24 hr at 4 °C.

Tissue was cut into 100 µm sections on a vibratome and stained for DAPI (4 µg/mL) to visualize cell nuclei. Confocal imaging of DiD and DAPI fluorescence confirmed that recordings in all mice were located in approximately the same subregions of the OFC.

Firing rate analysis, and identification of significantly discriminating or modulated cells

Spike sorting was performed using custom, semi-automated scripts written in MATLAB (Mathworks, Cambridge MA) for the identification of putative single units. The analysis combined all types of units (putative pyramidal cells and interneurons). The mean firing rate per unit was calculated by binning spike count data into 5 ms time steps, convolving with a Gaussian kernel (SD = 25 ms), and averaging across trials of the same stimulus type (either $CS1^+$, $CS1^-$, $CS2^+$, transfer). To determine whether a unit's activity significantly discriminated between $CS1^+$ and $CS1^-$ trials, we used a permutation test to detect significant differences in observed firing rate for each time step between these trials (Bakhurin et al., 2016). The firing rate was sampled from t = 0 to 1 s post CS1 onset in time steps of 5 ms. For each time step, the data from $CS1^+$ and $CS1^-$ trials were shuffled, and a new absolute difference in firing rate was calculated. This was repeated 10,000 times to obtain a distribution of permuted differences in firing rates. A unit was defined as being discriminating if the absolute value of the observed rate difference was higher than the 99^{th} percentile of the permuted distribution ($p = 0.01$). To calculate whether a unit's activity was significantly modulated we applied the same permutation analysis to compare cue-related firing with baseline activity. In each case, we used a 1 s period, corresponding to the duration of the cue, to determine cue-related firing, and compared this to a 4 s within-trial baseline period (-7 to -3 s, 4 s duration chosen to provide a smooth baseline average).

Onset, offset cell and population overlap analysis

Latency to peak firing during the period between the feature cue and CS (t = -2 to 0 s from CS onset) was estimated from the maximum average firing rate using 5 ms time bins and a Gaussian kernel convolution. Firing rate was calculated from the average of both $CS1^-$ and transfer trials (i.e., all trials containing a feature cue). The observed latency distribution across all recorded cells (Figure 4C) showed a good fit to the sum of two Lorentzian distributions. We defined the cutoff between onset and offset cells at the local minimum in the latency distribution, which occurred at t = -1.9 s from CS onset. The range of latency values was bounded from -2.5 to -1 s. To determine the overlapping population size predicted by chance between the feature, $CS1^-$ and $CS1^+$ cues, we first calculated the percentage of neurons per animal (n = 8) that was significantly modulated in response to these three individual cues. We then multiplied these three percentage values together to determine each animal's percentage of overlapping cells predicted by chance. This, in turn, was statistically compared to the observed overlap value of the corresponding animal using a paired t-test.

Network state prediction analysis

Analysis of cortical network state (Figures 5B, 5C) was performed separately for each animal, using both $CS1^-$ and transfer trials (i.e., all trials containing a feature cue). For the network state analysis, these two trial types were behaviorally indistinguishable because during the delay period, the animal had no prior knowledge of which CS it would subsequently receive. For each trial, the spike count for each unit was calculated for the 1 s period prior to the feature presentation (defined as the baseline, BL), and for the 1 s period occurring prior to the odor stimulus presentation (defined as the delay, DL). This resulted in two paired population rate vectors for each trial to be used in the classification algorithm. We used a binary support vector machine (SVM) classifier with a linear kernel, implemented in the LIBSVM library (version 3.21, (Chang and Lin, 2011)). The classifier was trained to distinguish between population rate

vectors on BL and DL periods (Figure 5B). We used a repeated five-fold cross-validation strategy, so that each training set contained four folds of trials, leaving the remaining fold for testing. Each fold of the data was used once for testing, ensuring that each trial was tested exactly once. During testing, each population rate vector in the tested fold was classified as belonging to either BL or DL periods. The classifier's performance was defined as the percentage of correctly classified BL and DL periods across all tested folds. We repeated this procedure 500 times, each time shuffling the order of trials allocated to the folds, to account for potential variability across trials in the population and to ensure the most accurate estimate of classifier performance. The average of all 500 accuracy scores was defined as the decoder accuracy score for each data set. To maximize decoder performance, we determined the optimal SVM misclassification cost parameter, *C*, via an iterative search across a range of parameters (also using five-fold cross-validation). The final value of *C* ranged from 0.002 to 0.0625. To determine the chance level of performance for each population, we shuffled the BL and DL labels on the data. We then applied the binary classifiers that were trained on observed data to the randomized datasets in a parallel cross-validation procedure. The mean decoder accuracy score on the randomized data (approximately 50%) was used as chance level for each data set.

We used a similar approach to classify whether delay period activity prior to incorrect $CS1^-$ trials, was more similar to the baseline period prior to correct $CS1^+$ trials, or the delay period prior to correct $CS1^-$ trials (Figures 5D, 5E). For each trial, the spike count for each unit was calculated for the 1 s baseline period prior to the odor presentation during correct $CS1^+$ trials (defined as the baseline prior to licking, BLL), the 1 s delay period occurring prior to the odor stimulus presentation during correct $CS1^-$ trials (defined as the delay prior to lick withholding, DLW), and the 1 s delay period occurring prior to the odor stimulus presentation during incorrect $CS1^-$ trials (defined as the delay prior to errant licking, DLL). This resulted in a population rate vector for each trial of each class to be used in the classification algorithm. Since there were an

uneven number of correct trial observations (unlike in the paired situation described for the BL versus DL activity classification) we equalized the numbers of correct trials by randomly subsampling the larger population down to the size of the smaller population. This ensured that classification would not be biased toward the type of trial that contained a greater numbers of observations. After training the classifier on balanced data from the two correctly performed trial types, we then tested all of the DLL observations on the model and asked whether the classifier was more likely to identify activity in the DLL period as a BLL or DLW period. We repeated this procedure 500 times, each time shuffling the order of trials prior to subsampling, thus creating a new classifier on new combinations of training trials. To maximize decoder performance, we determined the optimal SVM misclassification cost parameter, *C*. The optimal parameter for each dataset was determined by first subsampling from BLL and DLW trials, and performing five-fold cross validation decoding while systematically varying *C*. This procedure was performed 100 times, with each iteration containing a new combination of subsampled trials. Thus, we chose the *C* parameter that resulted in the highest BLL and DLW separation. The final values of *C* ranged from 0.001 to 0.125.

Experimental design and statistical analyses

All statistical tests were performed in MATLAB or Prism (GraphPad, La Jolla CA) software. The sample size, type of test used, and probability value is reported in the text and figure legends. All p values lower that 0.0001 are reported as $p < 0.0001$. One subject (animal # 1) was excluded from the analysis of Figure 5E for having only 1 DLL trial, which prevented a statistically sound analysis.

Results:

Behavioral responses reveal a negative occasion setting strategy

In the feature negative conditioning task, mice (n = 8) are exposed to conditioned odor stimuli

(CS1 and CS2, 1 s duration) that are either followed by reward if no feature cue (mild air puff) was present, or not followed by reward if a feature cue was present prior to the odor stimulus (Figure 1A). Therefore, the presence or absence of the feature cue determines the outcome on that trial. On training sessions, we presented three trial types with equal likelihood: $CS1^+$, $CS1^-$, and $CS2^+$ (Figure 1B, left). Thus, during this training period, the feature cue was presented in half of the CS1 trials, but never paired with the CS2 trials. The final training session, which coincided with electrophysiological recordings, included transfer trials in the form of the same feature cue followed by the CS2 cue (Figure 1B, right).

On the final training session, the percentage of $CS1^-$ trials with licking was significantly reduced relative to $CS1^+$ trials (Figures 1C, 1E; $p < 0.0001$, paired t-test). Thus, mice learned that the feature predicts an unrewarded outcome with respect to the CS1 cue. In order to determine the specificity of the feature-CS association, we introduced a small percentage (15%) of transfer trials, which animals encountered for the first time during the recording session. Animals showed a reduction in licking on transfer trials relative to $CS2^+$ trials (Figures 1D, 1F; $p = 0.03$, paired t-test). However, the inhibitory effect of the feature on licking in $CS1^-$ trials (62% median reduction, 28%, interquartile range, IQR) was significantly greater than its effect on transfer trials (9% median reduction, 21% IQR, $p < 0.0001$, paired t-test). Thus, the feature cue primarily suppressed CS1 elicited anticipatory licking behavior (compared with CS2), as predicted by the negative occasion setting model. This selectivity also suggests that information about the feature cue's presence is maintained during the delay period, in order to guide the animal's decision about whether to lick following the CS presentation.

Feature cues selectively inhibit OFC encoding of conditioned stimuli

Previous studies suggest that the OFC regulates feature negative behavior (Meyer and Bucci, 2016). However, the neural activity correlates of this behavior have not been studied in this brain area. We used silicon-based microprobes (4 silicon prongs with 64 electrodes each) to

simultaneously record from dozens of orbitofrontal units during the final training session (n = 8 mice, 48 to 119 single units per animal). After each recording, we verified the silicon prong locations using confocal microscopy (Figure 2A), and used these images to estimate the recording site and corresponding unit positions. We found that the measurements were primarily located in the ventral and lateral subregions of the OFC (Figure 2B).

Based on the finding that the feature cue predominantly diminished levels of anticipatory licking in response to the CS1, we hypothesized that the feature cue would modulate odor stimulus-evoked cortical activity. Consistent with this prediction, we observed that the presence of the feature, on $CS1^-$ trials, suppressed the OFC population's mean firing rate relative to $CS1^+$ trials during the CS presentation period (n = 585 units pooled across 8 mice, Figure 3A). We then separately examined the mean firing rate in each animal and found that the feature caused a significant reduction in firing rate during the 1 s CS1 presentation period (Figure 3D; $p = 0.016$, paired t-test). In contrast, we did not see any feature effect on mean CS2 evoked firing rate during transfer trials (Figures 3B, 3E; $p = 0.46$, paired t-test). Furthermore, we found a small but statistically significant difference ($p = 0.045$, paired t-test) between the feature-induced reduction in firing rate on CS1 compared to CS2 cues, demonstrating that the feature selectivity inhibits the encoding of the CS1 representation. We also found the OFC does not appear to encode choice, because we did not observe any difference in mean firing rate between $CS1^-$ trials with anticipatory licking and $CS1^-$ trials without licking (Figures 3C, 3F; $p = 0.30$, paired t-test).

To further examine the feature cue's effect on OFC neuronal responses to CS1 cues, we compared the firing rates between the $CS1^+$ and $CS1^-$ trials for each individual neuron during the 1 s cue presentation period (n = 585 units pooled across 8 mice). We found that a significant fraction of neurons had a lower firing rate in the $CS1^-$ trials (Figure 3G; $p < 0.0001$, paired t-test), suggesting that the feature suppressed the response of a large proportion of OFC neurons. We also found that the percentage of cells per animal that could discriminate between

380 $CS1^+$ and $CS1^-$ trials during the 1 s CS presentation period was significantly correlated with
381 behavioral discrimination (Figure 3H; n = 8 mice, Pearson r = 0.86, p = 0.012). Thus, the
382 greater the proportion of OFC units that distinguished between non-feature and feature trial
383 types, the better the animal was at correctly licking to $CS1^+$ trials and correctly withholding to
384 $CS1^-$ trials. Therefore, these electrophysiological measurements, together with the
385 corresponding behavioral tests, support the negative occasion setting model by showing that
386 the feature cue selectively suppresses OFC activity and anticipatory behavior following CS1
387 cues, but not CS2 cues.
388
389 **Temporally specific feature encoders have unique discriminatory properties**
390 To further understand the encoding properties of feature and CS1 cues, we examined the firing
391 patterns of individual neurons under different stimulus conditions. Across the recorded
392 population, we found that a large proportion of cells appeared to respond to individual cues
393 (feature, $CS1^+$, or $CS1^-$), or a combination of these cues (Figure 4A). To quantify this
394 relationship, we calculated the proportion of units that were significantly modulated by single
395 cues or different cue combinations. We found that across n = 8 mice, 53% (median, 20% IQR)
396 of the neurons responded to the feature whereas 51% (median, 15% IQR) and 42% (median,
397 28% IQR) of neurons responded to the CS1 in the $CS1^+$ and $CS1^-$ trials, respectively (Figure
398 4B). Notably, 31% (median, 19%, IQR) of the neurons responded to all three cues. This
399 overlap is significantly higher than chance levels (10%, 11% IQR), based on the total number of
400 identified units in the OFC (paired t-test, $p < 0.0001$), suggesting a common representation of
401 the cells that encode these stimuli. These results suggest that not only is OFC encoding of a
402 reward-associated stimulus (CS1) modulated by the feature cue, but that this circuit is strongly
403 tuned by stimuli that activate overlapping neuronal subpopulations.
404 In the population of feature responsive cells, we found evidence for heterogeneous
405 response properties, with some cells responding early, and others later to the feature cue

406 (Figure 4A). We calculated each unit's latency to peak firing during the feature period, and
407 found that the latency values appeared to cluster into two distinct firing groups (Figure 4C). One
408 group of neurons fired maximally around the feature onset time (onset cells), whereas another
409 group preferentially fired around the feature offset time (offset cells). We separately examined
410 the mean CS1-triggered firing rate of the onset and offset cells, and found that they appeared to
411 show different responses during the CS1 presentation period (Figure 4D). Specifically, the
412 mean firing rate of onset cells appeared markedly reduced in $CS1^-$ relative to $CS1^+$ trials (Figure
413 4D, top panel). This suggests that the CS1 representation associated with the onset population
414 is highly susceptible to suppressive properties of the feature. In contrast, the response of offset
415 cells to CS1 was less perturbed by the feature (Figure 4D, bottom panel). To quantify these
416 differences, we compared the number of cells within each group that significantly discriminated
417 between the $CS1^+$ and $CS1^-$ trial types during the 1 s CS1 presentation period. We found that
418 the onset group contained a significantly larger proportion of discriminating cells relative to the
419 offset group (Figure 4E; $p < 0.0001$, paired t-test).
420
421 **Feature cues trigger a distinct network state in the delay period**
422 If the feature cue influences subsequent OFC encoding of reward-conditioned stimuli, we
423 hypothesized that information about whether the feature cue was present is maintained in the
424 OFC throughout the delay period. Previous work suggests that delay periods in working
425 memory tasks often coincide with persistent firing patterns in the prefrontal cortex (Fuster and
426 Alexander, 1971; Goldman-Rakic, 1995; Fuster, 2005; Constantinidis, 2015). However, our
427 data revealed that the average firing rate in the OFC returns to baseline levels before the CS1
428 onset (Figures 3A, 4D), suggesting that the feature cue does not trigger persistent changes in
429 mean spiking activity. In support of this observation, there was no significant difference in mean
430 firing rate between the final 1 s of the delay period (DL), and a 1 s baseline period prior to
431 feature cue onset (BL, Figure 5A; $n = 8$ mice, $p = 0.11$, paired t-test). We therefore wondered if

the OFC could still maintain the information about the feature cue's presence during the delay period without any significant persistent activity signal. We speculated that if the OFC is maintaining this information, it does so through an 'activity silent' but distinct network state (Stokes et al., 2013; Stokes, 2015), that does not give rise to an overt change in mean firing rate. An alternative possibility is that another region outside the OFC is exclusively responsible for maintaining the feature cue information. To determine whether OFC networks exhibit dynamics during the delay period that are distinct from the baseline period, we used a decoder to distinguish between population activity in the DL and BL periods from the same trial containing a feature cue (Figure 5B). The decoder was applied to simultaneously recorded populations of cells from individual animals. Our results reveal that for all animals tested, the decoder performed significantly above chance levels in discriminating between activity in the BL and DL periods (Figure 5C; n = 8 mice, $p < 0.0001$, paired t-test). The average accuracy was 69 ± 2 % (mean ± SEM, dashed black line). To rule out any differential interaction between the paired BL and DL periods and the previous trial, we also compared the BL and DL periods from separate trials: $CS1^+$ and $CS1^-$, respectively. In this case, the decoder also performed significantly above chance levels in discriminating between activity in the BL and DL periods (n = 8 mice, $p < 0.0001$, paired t-test, data not shown). The average accuracy was 69 ± 3 % (mean ± SEM), which is very close to our value using the paired period method. A direct comparison revealed no significant differences (n = 8 mice, $p = 0.96$, paired t-test), indicating that both approaches produce the same result. Taken together, these findings suggest that, despite the absence of an overt change in mean population firing rate, the feature induces a distinct network state in the OFC during the delay period.

Finally, we examined whether OFC network dynamics during the delay period also provide information about the subsequent behavioral choice of the animal on that trial. In other words, is there a prospective code during the delay period that predicts whether or not the mouse will lick? To test this, we took advantage of the observation that mice sometimes licked incorrectly

458 during CS1$^-$ trials (Figure 1C). We trained a classifier to distinguish between population activity
459 occurring during the baseline period prior to correct CS1$^+$ trials (BLL), and the delay period prior
460 to correct CS1$^-$ trials (DLW). First, using cross-validation, we found that the classifier could
461 distinguish these periods above chance levels (n = 8 mice, $p < 0.0001$, paired t-test, data not
462 shown), consistent with a distinct network state during DL and BL periods shown in Figure 5C.
463 We next examined whether OFC population activity in the delay period prior to incorrect CS1$^-$
464 trials (DLL) was classified more frequently as a BLL or DLW period (Figure 5D). There was a
465 significant preference for the classifier to label DLL as a DLW period (Figure 5E; n = 7 mice, $p =$
466 0.018, paired t-test). The average accuracy was 64 ± 4 % (mean ± SEM, dashed black line).
467 Thus, it appears that in the OFC, the feature rather than the behavioral outcome (i.e., licking)
468 dictates the delay period network state. This is consistent with our earlier findings showing no
469 significant difference in mean firing rate between correct and incorrect CS1$^-$ trials (Figures 3C,
470 3F). These findings suggest that the feature triggers a network state that is maintained
471 throughout the delay period, which could function to downregulate the network's response to the
472 CS1 stimulus.
473
474 **Discussion:**
475 This is, to our knowledge, the first study to show the neural dynamics that may underlie an
476 occasion setter's ability to modulate behavior. A key insight that this study reveals is the
477 selective nature of the association between the feature cue and the conditioned odor stimulus
478 (Holland, 1984). The feature causes animals to suppress their conditioned responding in the
479 form of anticipatory licking to a trained stimulus (CS1). However, the ability of the feature to
480 suppress conditioned responding does not transfer to another stimulus (CS2) that had never
481 previously been paired with the feature. Neural recordings in the OFC complement this finding
482 by showing that the feature negatively modulates activity triggered by CS1, but this modulation
483 effect does not transfer to the CS2 cue. This lack of transfer, observed in both our behavioral

and neurophysiological data, rules out the simple Rescorla-Wagner model since this model posits that the feature's inhibitory properties should transfer to any CS paired with that reward. The fact that we did not observe transfer thus provides strong evidence against a direct inhibitory link between the feature cue and reward. Furthermore, our data suggest that the OFC may be involved in the task, because a measure of the level of OFC modulation by the feature (percent of cells per animal that discriminate between $CS1^+$ and $CS1^-$ trials) significantly correlates with an individual animal's behavioral discrimination. Together, our findings provide strong evidence for the negative occasion setting model (Holland, 1984; Lamarre and Holland, 1987) in which a feature cue can modulate the ability of a separate cue to retrieve its reward association.

Our data also suggest a possible OFC information transfer mechanism between the feature and conditioned odor stimulus during the delay period. Many studies on working memory have found persistent changes in mean population firing activity that accompany the delay period (Fuster and Alexander, 1971; Goldman-Rakic, 1995; Miller et al., 1996; Miller and Cohen, 2001; Pasternak and Greenlee, 2005; Liu et al., 2014). While we found that many cortical neurons were activated within ~1.5 s of the feature cue's presentation, this activity did not appear to persist into the final 1 s of the delay period, suggesting that the OFC subregions that were targeted here do not exhibit sustained changes in activity. Of course, this observation does not rule out the possibility that persistent activity occurs in other brain areas. On the other hand, a number of studies suggest that sustained activity is not necessary to retain task-relevant information (Jensen and Tesche, 2002; Howard et al., 2003; Riggall and Postle, 2012; Ester et al., 2015; Lundqvist et al., 2016). Intriguingly, an 'activity silent' model of working memory raises the possibility that information is retained in the patterns of network-level activity (Stokes et al., 2013; Stokes, 2015). To examine whether such an effect could be taking place in the OFC during the final 1 s of the delay period in our task, we used a machine learning-based decoding algorithm to assess whether this time period coincides with a distinct network state. In

all mice tested the decoder was able to accurately distinguish delay from baseline period activity at above chance levels, consistent with the activity silent working memory model (Stokes et al., 2013; Stokes, 2015). Thus, our data indicate that the OFC has the potential to transfer information about the feature cue across the delay period.

Our results suggest that the OFC uses the feature as a source of rule information in order to regulate behavioral responses. As discussed above, the degree to which the feature cue suppresses anticipatory licking correlates with its ability modulate neural activity to the conditioned odor stimulus. In contrast, we found no change in OFC activity during trials when animals incorrectly lick during a feature negative trial. Moreover, our classifier results suggest that the network state during the delay period prior to incorrect $CS1^-$ trials (DLL) is significantly different from the state during the baseline period prior to correct $CS1^+$ trials (BLL), even though both types of trials contain licking. These two pieces of evidence suggest that the OFC code is relatively insensitive to behavioral choice. Thus, our data are consistent with a number of other studies indicating the importance of rule encoding in the OFC (Buckley et al., 2009; Tsujimoto et al., 2009; 2012; Johnson et al., 2016; Sleezer et al., 2016).

The information coding properties revealed here provide insight into how the brain could quickly manipulate information at more abstract levels to regulate behavior. The feature appears to trigger a distinct network state that specifically interacts with its trained conditioned odor stimulus. This may occur by inducing a temporary functional reweighting of synaptic connections within OFC microcircuits (Fujisawa et al., 2008; Stokes, 2015). As a whole, this model fits well with the viewpoint that the OFC provides the animal with a cognitive map of task space (Roesch et al., 2006; Wilson et al., 2014; Cooch et al., 2015; Sharpe et al., 2015; Lopatina et al., 2016; Wikenheiser and Schoenbaum, 2016) because the extent to which the conditioned odor stimulus alters neural activity is mediated by the network state set by the feature. Taken together, our observations provide a potential mechanism that helps to explain

how animals can rapidly interpret the meaning of a conditionally rewarded cue to make timely behavioral decisions.

References:

Abramson CI, Cakmak I, Duell ME, Bates-Albers LM, Zuniga EM, Pendegraft L, Barnett A, Cowo CL, Warren JJ, Albritton-Ford AC, Barthell JF, Hranitz JM, Wells H (2013) Feature-positive and feature-negative learning in honey bees. J Exp Biol 216:224–229.

Bakhurin KI, Mac V, Golshani P, Masmanidis SC (2016) Temporal correlations among functionally specialized striatal neural ensembles in reward conditioned mice. Journal of Neurophysiology. 215:1521-32.

Buckley MJ, Mansouri FA, Hoda H, Mahboubi M, Browning PGF, Kwok SC, Phillips A, Tanaka K (2009) Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. Science 325:52–58.

Bueno JLO, Holland PC (2008) Occasion setting in Pavlovian ambiguous target discriminations. Behav Processes 79:132–147.

Chang C-C, Lin C-J (2011) LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2:27.

Constantinidis C (2015) Role of Prefrontal Persistent Activity in Working Memory. Front Syst Neurosci 9:181.

Cooch NK, Stalnaker TA, Wied HM, Bali-Chaudhary S, McDannald MA, Liu T-L, Schoenbaum G (2015) Orbitofrontal lesions eliminate signalling of biological significance in cue-responsive ventral striatal neurons. Nat Commun 6:7195.

557 Ester EF, Sprague TC, Serences JT (2015) Parietal and Frontal Cortex Encode Stimulus-
558 Specific Mnemonic Representations during Visual Working Memory. Neuron 87:893–905.

559 Franklin KBJ, Paxinos G (1997) The Mouse Brain in Stereotaxic Coordinates. Morgan
560 Kaufmann.

561 Fujisawa S, Amarasingham A, Harrison MT, Buzsáki G (2008) Behavior-dependent short-term
562 assembly dynamics in the medial prefrontal cortex. Nat Neurosci 11:823–833.

563 Fuster JM (2005) Memory. In: Cortex and Mind, pp 111–142. Oxford University Press.

564 Fuster JM, Alexander GE (1971) Neuron activity related to short-term memory. Science
565 173:652–654.

566 Goldman-Rakic PS (1995) Cellular basis of working memory. Neuron 14:477–485.

567 Guo ZV, Hires SA, Li N, O'Connor DH, Komiyama T, Ophir E, Huber D, Bonardi C, Morandell K,
568 Gutnisky D, Peron S, Xu NL, Cox J, Svoboda K (2014) Procedures for Behavioral
569 Experiments in Head-Fixed Mice. PLoS ONE 9(2): e88678.

570 Holland PC (1984) Differential effects of reinforcement of an inhibitory feature after serial and
571 simultaneous feature negative discrimination training. J Exp Psychol Anim Behav Process
572 10:461–475.

573 Holland PC (1989) Transfer of negative occasion setting and conditioned inhibition across
574 conditioned and unconditioned stimuli. J Exp Psychol Anim Behav Process 15:311–328.

575 Holland PC (1995a) Transfer of occasion setting across stimulus and response in operant
576 feature positive discriminations. Learning and Motivation 26:239–263.

577 Holland PC (1995b) The effects of intertrial and feature-target intervals on operant serial

feature-positive discrimination learning. Animal Learning & Behavior 23:411–428.

Holland PC (1997) Brain mechanisms for changes in processing of conditioned stimuli in Pavlovian conditioning: Implications for behavior theory. Animal Learning & Behavior 25:373–399.

Holland PC (2008) Cognitive versus stimulus-response theories of learning. Learn Behav 36:227–241.

Howard MW, Rizzuto DS, Caplan JB, Madsen JR, Lisman J, Aschenbrenner-Scheibe R, Schulze-Bonhage A, Kahana MJ (2003) Gamma oscillations correlate with working memory load in humans. Cereb Cortex 13:1369–1374.

Jensen O, Tesche CD (2002) Frontal theta activity in humans increases with memory load in a working memory task. Eur J Neurosci 15:1395–1399.

Johnson CM, Peckler H, Tai L-H, Wilbrecht L (2016) Rule learning enhances structural plasticity of long-range axons in frontal cortex. Nat Commun 7:10785.

Lamarre J, Holland PC (1987) Transfer of inhibition after serial feature negative discrimination training. Learning and Motivation. 18:319-342.

Liu D, Gu X, Zhu J, Zhang X, Han Z, Yan W, Cheng Q, Hao J, Fan H, Hou R, Chen Z, Chen Y, Li CT (2014) Medial prefrontal activity during delay period contributes to learning of a working memory task. Science 346:458–463.

Lopatina N, McDannald MA, Styer CV, Peterson JF, Sadacca BF, Cheer JF, Schoenbaum G (2016) Medial Orbitofrontal Neurons Preferentially Signal Cues Predicting Changes in Reward during Unblocking. J Neurosci 36:8416–8424.

599 Lundqvist M, Rose J, Herman P, Brincat SL, Buschman TJ, Miller EK (2016) Gamma and Beta
600 Bursts Underlie Working Memory. Neuron 90:152–164.

601 Meyer HC, Bucci DJ (2016) Imbalanced Activity in the Orbitofrontal Cortex and Nucleus
602 Accumbens Impairs Behavioral Inhibition. Curr Biol. 26:2834:2839.

603 Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev
604 Neurosci 24:167–202.

605 Miller EK, Erickson CA, Desimone R (1996) Neural mechanisms of visual working memory in
606 prefrontal cortex of the macaque. J Neurosci 16:5154–5167.

607 Nallan GB, Brown MB, Edmonds C, Gillham V, Kowalewski K, Miller JS (1981) Transfer effects
608 in feature-positive and feature-negative learning by adult humans. Am J Psychol 94:417–
609 429.

610 Pace GM, McCoy DF (1981) Effects of stimulus contact on the feature-positive effect. Am J
611 Psychol 94:153–158.

612 Pace GM, McCoy DF, Nallan GB (1980) Feature-positive and feature-negative learning in the
613 rhesus monkey and pigeon. Am J Psychol 93:409–427.

614 Pasternak T, Greenlee MW (2005) Working memory in primate sensory systems. Nat Rev
615 Neurosci 6:97–107.

616 Rescorla RA (1969) Pavlovian conditioned inhibition. Psychol Bull 72:77–94.

617 Rescorla RA, Holland PC (1977) Associations in Pavlovian conditioned inhibition. Learning and
618 Motivation 8:429–447.

619 Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: Variations in the

620 effectiveness of reinforcement and nonreinforcement. In: Classical conditioning II: Current
621 research and theory, pp 64-99. Meredith Corporation.

622 Riggall AC, Postle BR (2012) The relationship between working memory storage and elevated
623 activity as measured with functional magnetic resonance imaging. J Neurosci 32:12990–
624 12998.

625 Robinson S, Keene CS, Iaccarino HF, Duan D, Bucci DJ (2011) Involvement of retrosplenial
626 cortex in forming associations between multiple sensory stimuli. Behav Neurosci 125:578–
627 587.

628 Roesch MR, Taylor AR, Schoenbaum G (2006) Encoding of time-discounted rewards in
629 orbitofrontal cortex is independent of value representation. Neuron 51:509–520.

630 Schmajuk NAE, Holland PCE (1998) Occasion setting: Associative learning and cognition in
631 animals. (Schmajuk NA, Holland PC, eds). Washington: American Psychological
632 Association.

633 Sharpe MJ, Wikenheiser AM, Niv Y, Schoenbaum G (2015) The State of the Orbitofrontal
634 Cortex. Neuron 88:1075–1077.

635 Shobe JL, Claar LD, Parhami S, Bakhurin KI, Masmanidis SC (2015) Brain activity mapping at
636 multiple scales with silicon microprobes containing 1,024 electrodes. Journal of
637 Neurophysiology 114:2043–2052.

638 Sleezer BJ, Castagno MD, Hayden BY (2016) Rule Encoding in Orbitofrontal Cortex and
639 Striatum Guides Selection. J Neurosci 36:11223–11237.

640 Stokes MG (2015) “Activity-silent” working memory in prefrontal cortex: a dynamic coding
641 framework. Trends Cogn Sci (Regul Ed) 19:394–405.

642 Stokes MG, Kusunoki M, Sigala N, Nili H, Gaffan D, Duncan J (2013) Dynamic coding for
643 cognitive control in prefrontal cortex. Neuron 78:364–375.

644 Tsujimoto S, Genovesio A, Wise SP (2009) Monkey orbitofrontal cortex encodes response
645 choices near feedback time. J Neurosci 29:2569–2574.

646 Tsujimoto S, Genovesio A, Wise SP (2012) Neuronal activity during a cued strategy task:
647 comparison of dorsolateral, orbital, and polar prefrontal cortex. J Neurosci 32:11017–11031.

648 Wikenheiser AM, Schoenbaum G (2016) Over the river, through the woods: cognitive maps in
649 the hippocampus and orbitofrontal cortex. Nat Rev Neurosci 17:513–523.

650 Wilson RC, Takahashi YK, Schoenbaum G, Niv Y (2014) Orbitofrontal Cortex as a Cognitive
651 Map of Task Space. Neuron 81:267–279.

652
653
654
655
656
657
658
659
660
661
662
663
664

Figure Legends:

Figure 1. Distinct associations form from feature negative conditioning.

(A) Schematic of the four distinct trial types used during training and recording sessions. In rewarded trials ($CS1^+$ and $CS2^+$), different conditioned odor stimuli (CS1 or CS2, 1 s duration) predicted the delivery of reward. In unrewarded trials ($CS1^-$ and transfer trials), when the same odor stimuli were preceded by a feature cue (mild air puff, 0.5 s duration), there was no reward. Orange bar denotes the feature, grey bar denotes CS1, green bar denotes CS2, black bar denotes reward. **(B)** Probability of presenting each trial type during initial training (left) and on the final training session corresponding to recording (right). All behavioral and electrophysiological results are from the final day. **(C & D)** Average lick rate as a function of time during all rewarded and unrewarded trials. Dashed lines represent the onset and offset times of the indicated cue. Data represent mean ± SEM (n = 8 mice). Grey bar: CS1, green bar: CS2, orange bar: feature. **(E)** The feature significantly reduces the likelihood that animals express anticipatory licking (t = 0 to 2.5 from odor onset) in CS1 trials ($p < 0.0001$, paired t-test). **(F)** The feature significantly suppresses the likelihood of anticipatory licking in transfer trials ($p = 0.03$, paired t-test).

Figure 2. Silicon microprobe recordings in the OFC.

(A) Representative confocal image of a coronal section showing the recording position of the silicon microprobe containing 4 prongs. Prior to insertion, the prongs were painted with DiD (red) to facilitate visualization. The section was stained with DAPI (blue). **(B)** Coronal section from the Franklin and Paxinos mouse brain atlas (2.35 mm anterior to bregma (Franklin and Paxinos, 1997)) annotated with the estimated position of each putative unit (red dot) in relation to the OFC structure.

Figure 3. Cue-dependent modulation of OFC activity.

(A, B & C). Mean firing rate as a function of time in different trial types. Dashed lines represent the onset and offset times of the indicated cue. Data represent mean ± SEM (n = 585 units). Grey bar: CS1, green bar: CS2, orange bar: feature. **(A)** Comparison of $CS1^+$ with $CS1^-$ trials. **(B)** Comparison of $CS2^+$ with transfer trials. **(C)** Comparison of $CS1^-$ trials with licking or without anticipatory licking. **(D, E & F**) Mean firing rate per animal during the CS presentation period (t = 0 to 1 s), in different trial conditions. Data represent individual animals (n = 8). **(D)** $CS1^-$ trials exhibit significantly lower firing than $CS1^+$ trials ($p = 0.016$, paired t-test). **(E)** There is no significant difference in mean firing between transfer trials and $CS2^+$ trials ($p = 0.46$, paired t-test). **(F)** There is no significant difference in mean firing between $CS1^-$ trials with licking and those without licking ($p = 0.3$, paired t-test). **(G)** Comparison of the average firing rate per unit during the CS cue presentation period (t = 0 to 1 s) between $CS1^+$ and $CS1^-$ trial types. Across the population (n = 585) there was a significant bias toward lower firing during $CS1^-$ trials ($p < 0.0001$, paired t-test). **(H)** Behavioral discrimination (percent correct $CS1^+$ trials minus percent incorrect $CS1^-$ trials) is significantly correlated with the percentage of OFC units per animal that discriminate between $CS1^+$ and $CS1^-$ trials (Pearson $r = 0.82$, $p = 0.012$).

Figure 4. Identification of temporally distinct feature encoding populations.

(A) Mean normalized firing rate as a function of time of the recorded population (n = 585 cells). Each cell's firing rate is normalized to its peak firing rate on CS^- trials (top panel) and CS^+ trials (bottom panel). Units are ordered by latency to peak firing relative to onset of the feature cue (FT). Units are plotted in the same order in the top and bottom panels (red indicates high firing rate). **(B)** Venn diagram showing the overlapping relationship between units that were significantly modulated by the feature (orange), the CS1 cue during $CS1^+$ trials (magenta), and the CS1 cue during $CS1^-$ trials (blue). Values represent the median percentage of modulated cells across n = 8 animals. **(C)** Distribution of the latency to peak firing for the recorded

717 population (n = 585 units). Two major peaks were resolved using Lorentzian curve fits (red
718 line). Dashed orange lines demarcate the onset and offset cell populations. **(D)** Mean firing
719 rate as a function of time during $CS1^+$ and $CS1^-$ trials. The top and bottom panels are
720 comprised of onset and offset cells, respectively. Data represent mean ± SEM (n = 585 units).
721 The orange shaded area represents the time during the feature cue presentation. **(E)** The
722 percentage of cells that discriminated between $CS1^+$ and $CS1^-$ trials was significantly higher in
723 the onset cell population (n = 8 mice, $p < 0.0001$, paired t-test).
724
725 **Figure 5. A distinct network state initiated by the feature cue.**
726 **(A)** There is no significant difference in mean OFC firing rate during the final 1 s of the delay
727 period (DL) and a 1 s baseline period prior to feature cue presentation (BL, $p = 0.11$, paired t-
728 test). **(B)** Strategy used to determine whether the network state in the BL period is distinct from
729 that of the DL period. This two-step process required training (top dashed box) and testing
730 (bottom dashed box) a binary classifier. During testing, each period (BL, green arrows and DL,
731 blue arrows) was classified as either a correct match (e.g., BL classified as BL, solid arrow) or
732 an incorrect match (e.g., BL classified as DL, dashed arrow). **(C)** Mean classifier accuracy per
733 animal of the classifier in B (accuracy defined as the percentage of correctly classified BL and
734 DL periods across all tested folds, black) was significantly above chance levels shown in red (n
735 = 8, $p < 0.0001$, paired t-test). The average accuracy across the experimental group was 69 ± 2
736 % (mean ± SEM, dashed black line). **(D)** Strategy used to classify whether delay period activity
737 prior to incorrect $CS1^-$ trials (DLL), was more similar to the baseline period prior to correct $CS1^+$
738 trials (BLL), or the delay period prior to correct $CS1^-$ trials (DLW). The classifier was trained (top
739 dashed box) to distinguish population activity during BLL periods from DLW periods. During
740 testing (bottom dashed box) DLL activity was compared to BLL and DLW activity and classified
741 as more similar to either BLL (dashed line) or DLW (solid line). **(E)** Mean classifier accuracy per
742 animal of the classifier in D (accuracy defined as the percentage of DLL periods that were

743 labeled as DLW, black) was significantly above chance levels shown in red (n = 7, p = 0.018,
744 paired t-test). The average accuracy across the experimental group was 64 ± 4 % (mean ±
745 SEM, dashed black line). Note that animal # 1 only had 1 DLL trial and was excluded from the
746 analysis in E. Error bars in C and E represent 95% confidence intervals across all iterations and
747 dashed lines represent the average values across all animals.