The intersection between the representation of the stimuli and the choice by neural ensembles in the primary visual cortex of the macaque.

Veronika Koren [1,2,*], Ariana R. Andrei [3], Ming Hu[4], Valentin Dragoi [3], Klaus Obermayer [1,2]

1 Neural Information Processing Group, Institute of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, Berlin, 10587, Germany **2** Bernstein Center for Computational Neuroscience Berlin, Germany **3** Department of Neurobiology and Anatomy, University of Texas Medical School, Houston, Texas, 77030, US **4** Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, US

[*] Correspondence: veronika.koren@ni.tu-berlin.de

Highlights

- Learning from both stimuli and behavioral choice generalizes to the representation of the choice alone.
- With generalization of learning, the choice signal can be read-out from parallel spike trains in V1.
- Correct attribution to binary coding pools and correct spike timing are necessary for the read-out.
- Bursty neurons convey more information than non-bursty neurons. Discrimination is the strongest in the superficial layer of the cortex.

Abstract

Representation of stimuli by neural ensembles and the correlation of neural activity with the behavioral choice are, in principle, two different computational problems, however, it is only the intersection between the two that is relevant for animal's behavior. Here, we propose a decoding model that learns its decoding weights in the presence of the information on both the stimulus class and the future behavioral choice. We then test the model on trials that only differ in the choice and show that the choice can be read-out from the activity of populations in V1. The read-out model is a linear weighted sum of spikes, decoding the behavioral choice in single trials and without temporal averaging. The generalization of learning suggests that the representation of the stimulus class and of the behavioral choice have a non-zero intersection. In addition, we show how the spike timing is required for discrimination, that bursty neurons carry more information than non-bursty neurons and that neurons in the superficial layer are the most important for discrimination.

Keywords: Visual cortex; parallel spike trains; decoding; stimulus; choice; computation; representation; neural network; population code; information; discrimination

Introduction

In real-world scenarios, animals frequently take decisions based on the observation of their environment. In the context of such decision-making, it has been understood that encoding of sensory information and utilizing this information for guiding the behavior do not necessarily overlap [26]. The information about the stimuli might be represented in the neuronal activity, without being read-out by downstream neurons. If the information is not read-out, it does not have any causal relation to behavior and is in fact irrelevant for behavior. It has been suggested that the intersection between the represented information and the behavior can only be understood through intervention-the experimenter must understand the representation, perturb it, and observe the effect of the perturbation on the behavior [26]. If represented information does not have an intersection with the behavior, perturbing it will leave the behavior intact, and if a particular representation has an intersection, perturbation will change the behavior. Unfortunately, such a setting is difficult to realize in experiments and is only starting to be utilized *in vivo* [16].

Here, we address the intersection between the representation of stimuli and the representation of choice indirectly, by studying the transfer of learning from the classification problem in the presence of the information on both the stimulus class and the behavioral choice to the read-out of the behavioral choice. The animal subject visualizes pairs of stimuli that can either be matching or non-matching and decides whether the two stimuli were same or different (table 1). The choice of the animal is either compatible with the visualized stimulus class (correct behavior), or not (incorrect behavior). We consider two classification problems, one where classification relies on the information on both the stimuli and the choice (*stimulus + choice*), and another, where classification relies on the information about the choice alone (*choice*, table 1). We ask, does learning in the context of *stimulus + choice* generalize to representation in the context of *choice*? To address this question, we train an optimal linear decoder in the context of *stimulus + choice* and use the decoder to construct the representation in the context of *choice*.

Stimulus	Choice	Performance	Info. content
non-match	"different"	correct	stimulus+choice
match	"same"	correct	
non-match	"different"	correct	choice
non-match	"same"	incorrect	

Table 1. Informational content of two classification problems.

If the representation in the context of *stimulus + choice* mainly relies on the difference of stimuli, and if signals related to the stimuli and the choice are independent, learning in context of *stimulus + choice* cannot be informative for the representation in the context of *choice*. If, on the contrary, classification in the context of *stimulus + choice* at least partially depends on a purely choice-related information, and/or the signals related to stimuli and the choice are not independent, learning in the context of *stimulus + choice* could transfer to the representation of the *choice*, showing a non-zero intersection between the two classification problems.

Materials and Methods

Methods availability

Code and dataset will be freely available in a pubic GitHub repository.

Experimental model and subject details

Ethics statement All experiments were conducted in accordance with protocols approved by The Animal Welfare Committee (AWC) and the Institutional Animal Care and Use Committee (IACUC) for McGovern Medical School at The University of Texas Health Science Center at Houston (UTHealth), and met or exceeded the standards proposed by the National Institutes of Healths Guide for the Care and Use of Laboratory Animals.

Animal subjects Two male rhesus macaques (Macaca mulatta; M1, 7 years old, 15kg; M2, 11 years old, 13kg) were used in this study. Subjects were housed individually (after failed attempts to pair house) in cages sized 73 x 69 x 31 or 73 x 34.5 x31 inches, in close proximity to monkeys in adjacent cages, allowing for visual, olfactory and auditory contact. Toys were given in rotation, along with various puzzles, movies and radio programming as environmental enrichments. Monkeys were fed a standard monkey biscuit diet (LabDiet), that was supplemented daily with a variety of fruits and vegetables. Subjects had been previously trained to perform visual discrimination task, and each implanted with a titanium head post device and two 19mm recording chambers (Crist Instruments) over V1 and V4. All surgeries were performed aseptically, under general anesthesia maintained and monitored by the veterinary staff from the Center for Laboratory Animal Medicine and Care (CLAMC), with appropriate analgesics as directed by the specialized non-human primate veterinarian at CLAMC. During the study the animals had unrestricted access to fluid, except on days when behavioral tasks were performed. These days, animals had unlimited access to fluid during the behavioral task, receiving fluid for each correctly completed trial. Following the behavioral task, animals were returned to their home cage and were given additional access to fluid. The minimal daily fluid allotment was 50ml/kg (monkeys were weighed weekly), though monkeys could drink more through their participation in the task. During the study, the animals health and welfare was monitored daily by the veterinarians and the animal facility staff at CLAMC and the labs scientists, all specialized with working with non-human primates.

Method details

Experimental setup Animals performed a delayed match-to-sample task on visual stimuli. The trial started after 300 ms of successful fixation within the fixation area and consisted in displaying the target and the test stimuli, with a delay period in between. The target and the test stimuli were either identical (condition "match") or else the test stimulus was rotated with respect to the target stimulus (condition "non-match"). The target and the test stimuli were shown for 300 ms each while the delay period had a random duration between 800 and 1000 ms. The task of the animal was to decide about the similarity of the target and the test stimuli by holding a bar for "different" and releasing the bar for "same". Stimuli were complex naturalistic images in black and white,showing an outdoor scene. The subject was required to respond within 200 and 1200 ms with respect to the offset of the test stimulus, otherwise the trial was discarded. The difference in orientation of the test stimulus ranged between 3 and 10 degrees and was calibrated on-line in order to have on average 70 percent correct responses on non-matching stimuli.

Recordings Recording were made with laminar electrodes, measuring the multi-unit signal with 16 recording channels (0.1 mm spacing between adjacent contacts). Electrodes were inserted perpendicularly to the cortical surface, and calibrated such that neurons from the two areas had overlapping receptive fields. In part of sessions, recordings were made in V1 and V4 simultaneously, with one laminar electrode in each area, while in other sessions, only V1 has been recorded. Since recording sessions were performed on different days, it is extremely unlikely that the electrode captured same neurons in different recording sessions. We therefore treated spike-sorted units from different recording sessions as distinct cells. The analysis included all cells that responded to the stimulus with a 4-fold increase of the firing rate with respect to the baseline.

The data was collected in 20 recording sessions in V1, which gave 160 neurons. We analyzed 3 conditions, "correct match" (CM), "correct non-match" (CNM) and "incorrect non-match" (INM), where "correct/incorrect" refers to the behavioral performance and "match/non-match" refers to the stimulus class. In

condition "incorrect match", there were not enough trials to perform the analysis. Below, table 2 reports the summary statistics on the number of trials in each condition.

Condition	Nb. trials min.	Nb. trials max.	Nb trials average
CM	51	290	118
CNM	42	230	107
INM	21	98	39

Table 2. The minimal, maximal, and average number of trials across recording sessions, for each condition.

Quantification and statistical analysis

The analysis was done with Matlab, Mathworks, version R2017b. The spike train of a single neuron n in trial j is defined as a binary vector of zeros and ones,

$$o_{n,j}(t_k) = \begin{cases} 1, & \text{if neuron } n \text{ in trial } j \text{ spikes during the } k\text{-th millisecond} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $n = 1, ..., N$ is the neural index, $k = 1, ..., K$, is the time index with step of 1 millisecond, and $j = 1, .., J_1, J_1 + 1, ..., J_2, J_2 + 1, ...J$ is the trial index. Trials were collected in conditions CM ($j = 1, ..., J_1$), CNM ($j = J_1 + 1, ..., J_2$) and INM ($j = J_2 + 1, ..., J$).

The coefficient of variation for the neuron n in trial j is defined as follows,

$$CV2_{n,j} = 2\frac{|ISI_{i+1} - ISI_i|}{ISI_{i+1} + ISI_i}$$

where ISI is the inter-spike interval with index $i = 1, ..., N_{int}$. We report trial-averaged results, distinguishing trials from conditions CNM (decision "same") and INM (decision "different").

$$\begin{aligned} CV2_n^{\text{different}} &= \frac{1}{J_2 - J_1} \sum_{j=J_1+1}^{J_2} CV2_{n,j} \\ CV2_n^{\text{same}} &= \frac{1}{J - J_2} \sum_{j=J_2+1}^{J} CV2_{n,j} \end{aligned} \tag{2}$$

Learning of structural features from parallel spike counts

We compute spike counts in the time window of $[0, K]$ ms with respect to the onset of the stimulus (target or test), $s_{n,j} = \sum_{t_k=1}^{K} o_{n,j}(t_k)$. Spike counts are z-scored:

$$\tilde{s}_{n,j} = \frac{s_{n,j} - \langle s_{n,j} \rangle_j}{\sqrt{\text{Var}_j\ (s_{n,j})}}, \tag{3}$$

where $\langle s_{n,j} \rangle_j$ is the empirical mean and $\text{Var}_j\ (s_{n,j})$ is the empirical variance across trials from all conditions.

The decoding weights are learned using conditions CM and CNM, which differ in both stimuli and choice behavior. We use the linear Support Vector Machine (SVM) on parallel spike counts, and compute the vector of feature weights of the classifier. These weights extract the structure of parallel spike counts. We then use this structure to reconstruct the choice signal in conditions CNM and INM, which only differ in the choice behavior. The choice signal is computed from parallel spike trains (without averaging across time) and in single trials.

The training set comprises all trials from condition CM and half of the trials from condition CNM (trials with index $j = 1, ..., J_1, J_1 + 1, ..., \frac{J_2 - J_1}{2}$). The reconstruction of the choice signal is computed on a hold-out set, utilizing the remaining half of the trials from condition CNM and all trials from condition INM (trials

with index $j = \frac{J_2-J_1}{2} + 1, ..., J_2, J_2+1, ..., J$). The split of trials in training and reconstruction set in condition CNM is cross-validated with Monte-Carlo method, using N_{cv} random splits. The split of trials in training and reconstruction set is non-overlapping, such that no trials that have been used for training appear in the reconstruction set. All reported results are averaged across cross-validations.

Estimation of the population vector In the N-dimensional space of inputs, one sample is the vector of z-scored spike counts of N simultaneously recorded neurons in trial j, $\mathbf{s}_j = [\tilde{s}_{1,j}, \tilde{s}_{2,j}, ..., \tilde{s}_{N,j}]^T$. Linear SVM searches for an $N-1$-dimensional plane (a hyperplane) that optimally separates points in conditions CNM and CM. The hyperplane is defined as follows,

$$H_0 : \mathbf{w}^T\mathbf{s}_j + b = w_1 s_{1,j} + w_2 s_{2,j} + ... + w_N s_{N,j} + b = 0 \tag{4}$$

where $\mathbf{w}$ is the vector of feature weights and b is the offset of the hyperplane from the origin. On each side of H_0, we can define a hyperplane that verifies the following:

$$\begin{aligned} H_1 : \mathbf{w}^T\mathbf{s}_j + b = -1 \\ H_2 : \mathbf{w}^T\mathbf{s}_j + b = 1 \end{aligned} \tag{5}$$

If the problem is linearly separable, all training samples verify the following the inequality,

$$\begin{aligned} \mathbf{w}^T\mathbf{s}_j + b \leq -1 \text{ if } y_j = -1 \\ \mathbf{w}^T\mathbf{s}_j + b \geq 1 \text{ if } y_j = 1 \end{aligned} \tag{6}$$

where $y_j \in \{-1, 1\}$ is the class label ($y_j = -1$ in condition CNM and $y_j = 1$ in condition CM). Training the linear SVM consists in maximizing the number of correctly classified samples and, at the same time, minimizing the distance between H_1 and H_2, which can be expressed with the Lagrangian.

$$L_p = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{j=1}^{\frac{J_2-J_1}{2}} \lambda_j[y_j(\mathbf{w}^T\mathbf{s}_j + b) - 1] \tag{7}$$

The first term on the right hand side ensures the maximal distance between hyperplanes H_1 and H_2, and the second term on the right ensures correct classification. As the derivative of the Lagrangian with respect to $\mathbf{w}$ is set to zero, we get the expression for the vector of weights.

$$\mathbf{w} = \sum_{j=1}^{\frac{J_2-J_1}{2}} \lambda_j y_j \mathbf{s}_j \tag{8}$$

Since $\lambda_j \neq 0$ only for trials that define the margin (points that lie on $H1$ or on $H2$), the weight vector only depends on support vectors,

$$\mathbf{w} = \sum_{q=1}^{Q} \lambda_q y_q \mathbf{v}_q \tag{9}$$

where $q = 1, 2, ..., Q$ are the support vectors, with $Q < \frac{J_2-J_1}{2}$. The weight vector is normalized, $\tilde{\mathbf{w}} = \mathbf{w}/||\mathbf{w}||$, with $||\mathbf{w}|| = (w_1^2 + ... + w_N^2)^{\frac{1}{2}}$. We refer to the vector $\tilde{\mathbf{w}} = [\tilde{w}_1, \tilde{w}_2, ..., \tilde{w}_n, ..., \tilde{w}_N]^T$ as the population vector and to the $n-$th entry of $\tilde{\mathbf{w}}$, $\tilde{w}_n$, as the decoding weight of the neuron n.

Representation of the choice signal as a read-out of parallel spike trains

Traditionally, learned features are applied on a hold-out set, using the same input statistics and the same input classes (or conditions) as the ones used for learning. Such a decoding model has been studied in our previous work, where we train and test the decoder on spike counts from conditions CNM and CM [15]. Further on,

we tested a decoding model that learns structural features on spike counts, and then applies them on spike trains [14]. In the latter, learning and reconstruction steps utilize the same conditions (CNM and CM), but different input statistics (spike counts for learning and spike trains for reconstruction). In the present work, we utilize the same decoding model as [14], but apply learned features on a different classification problem. We use conditions CM and CNM in the learning step, and conditions CNM and INM in the reconstruction step. In this sense, we are testing the generalization of learning. In the following, we describe the reconstruction part. For further information about the method, see [14].

Consider the vector of spike trains of N simultaneously recorded neurons.

$$\mathbf{o}_j(t_k) = [o_{1,j}(t_k), o_{2,j}(t_k)..., o_{N,j}(t_k)]^T \tag{10}$$

The population signal is a projection of the vector of spike trains on the vector of weights,

$$x_j(t_k) = F(\tilde{\mathbf{w}}^T \mathbf{o}_j(t_k)) - \langle F(\tilde{\mathbf{w}}^T \mathbf{o}_j(t_k))\rangle_j \tag{11}$$

where $\langle F(\tilde{\mathbf{w}}^T \mathbf{o}_j(t_k))\rangle_j$ is the trial average, utilizing all trials of the reconstruction set. By subtracting the trial average, we compute the deviation of the signal from the mean. We argue that the deviation of the signal from its mean, rather than the absolute value of the signal, might be biologically relevant. As the transfer function, $F(y(t_k))$, we use a convolution with an exponential kernel,

$$F(y) = \sum_{\tau \in T} y(t_k - \tau) u(\tau) \tag{12}$$

with $u(\tau) = exp(-\lambda\tau),\ \tau \in T$, with support $T = \{-100, ..., 100\}$ ms. Convolution with an exponential kernel models the causal effect of the presynaptic spike on the neural membrane of the read-out neuron. Notice that $x_j(t_k)$ is a time-resolved, low-dimensional representation of parallel spike trains in single trials.

To test the discriminability of conditions CNM and INM, we average the population signal across trials, distinguishing conditions CNM (decision "different") and INM (decision "same"),

$$\begin{aligned} x^{\text{same}}(t_k) &= \frac{2}{J_2 - J_1} \sum_{j=J'}^{J_2} x_j(t_k) \\ x^{\text{different}}(t_k) &= \frac{1}{J - J_2} \sum_{j=J_2+1}^{J} x_j(t_k) \end{aligned} \tag{13}$$

where $J' = \lceil \frac{1}{2}(J_1 + J_2) \rceil + 1$, and $\lceil z \rceil$ is the ceiling function . The significance of the difference between the population signal for the decision "same" and "different" is evaluated with the permutation test. We compute the difference of the population signal in every recording session,

$$\Delta x(t_k) = x^{\text{same}}(t_k) - x^{\text{different}}(t_k) \tag{14}$$

and average across sessions. We then rank $\Delta x(t_k)$ among $\Delta x_p^{perm}(t_k)$, where the latter has been computed with random weights, and the class labels in the validation set have been randomly permuted. Random weights were drawn from the uniform distribution with the same range as the regular weights. The permutation procedure is repeated N_{nperm}-times and gives a distribution of results for each time step. When the result of the true model, $\Delta x(t_k)$, appears outside of the distribution of results of the null model, $\Delta x_p^{perm}(t_k)$ for $p = 1, 2, ..., N_{perm}$, we consider that signals $x^{\text{same}}(t_k)$ and $x^{\text{diff}}(t_k)$ have been successfully discriminated.

Criteria for division into subpopulations

Sign of the weight We separate the population with respect to the sign of the decoding weight, distinguishing neurons with positive weight ($\tilde{w}_n > 0$, *plus* neurons) and negative weight ($\tilde{w}_n < 0$, *minus* neurons).

Informativeness We distinguish informative and uninformative neurons by ranking the absolute value of the weight, $|\tilde{w}_n|$, among the distribution of weights of models with permuted class labels. The strength of the weight of a single neuron n is a scalar, and the same result for models with permuted class labels is a distribution of N_{perm} values, where N_{perm} is the number of random permutations of class labels. If the strength of the weight of the neuron n is ranked within the first 25 % of weights from the null model, we assume that the neuron n is informative, and assume it is uninformative otherwise.

Burstiness We distinguish bursty and non-bursty neurons with a simple method based on the power spectrum of spike trains. It has been shown that bursty neurons have decreased power spectrum in low and middle frequency ranges [6]. We compute the power spectrum of spike trains for every single neuron, using multiplication of the spike train with Slepian tapers to increase the reliability of the estimation [27]. We use 5 Slepian tapers and the time window of $K = 400$ ms. Power spectra are normalized with neuron's firing rate. As a reference, homogeneous Poisson process has a flat power spectrum of 1, and typically, the power spectrum of a bursty neuron is lower than 1 for low and middle frequency ranges. We compute the normalized sum of the power spectrum for frequencies between 10 and 200 Hz. Frequencies below 10 Hz are discarded, since the power spectrum at those frequencies cannot be accurately estimated due to the short time window K. Significance of the sum under the power spectrum is estimated with the permutation test. The null model is computed by randomly permuting, without repetition, the time index of spike trains $N_{perm}-$ times. We rank the sum under the power spectrum of the regular model among the same results for models with permuted time index. We assume that the neuron n is bursty if the normalized sum under its power spectrum is significantly below the same result of models with permuted spike timing ($\alpha = 0.05$). If the neuron does not fulfill this criterion, it is assumed to be non-bursty.

Cortical layers We distinguish three cortical layers, the superficial (supragranular, SG), the middle (granular, G) and the deep layer (infragranular, IG, [10]). The method for determining cortical layers utilizes the covariance matrix of the current source density (see methods and fig. 6 in [14],).

Population signal in subpopulations

Sign-specific population signal is computed by removing the information from the spike train of neurons with the opposite sign. We remove the information by randomly permuting class labels "same" and "different" in the reconstruction step. When the class label in the reconstruction step is permuted, we get a random association between the weight and the class label of the spike train, resulting in a signal that is close to zero at all times. As an example, the population signal of *plus* neurons is computed with the spike train $\mathbf{o}_{j,p}^{perm,+}(t_k)$, where the label of the spike train is correct for *plus* neurons, and random (i.e., correct or incorrect with equal probability) for *minus* neurons.

$$x_{j,p}^{+}(t_k) = F(\tilde{\mathbf{w}}^T \mathbf{o}_j^{perm,+}(t_k)) - \langle F(\tilde{\mathbf{w}}^T \mathbf{o}_j^{perm,+}(t_k)) \rangle_j \tag{15}$$

Similarly, the signal of *minus* neurons is computed by utilizing the spike train $\mathbf{o}_{j,p}^{perm,-}(t_k)$, where the label of the spike train is correct for *minus* neurons, and random for *plus* neurons.

$$x_{j,p}^{-}(t_k) = F(\tilde{\mathbf{w}}^T \mathbf{o}_j^{perm,-}(t_k)) - \langle F(\tilde{\mathbf{w}}^T \mathbf{o}_j^{perm,-}(t_k)) \rangle_j \tag{16}$$

Random permutation is repeated $N_{perm}-$times, with $p = 1, 2, ..., N_{perm}$ random permutations, without repetition, of the order of trials. As before, we average each of the signals across trials, distinguishing conditions

"same" and "different".

$$x_p^{+,\text{same}}(t_k) = \frac{2}{J_2 - J_1} \sum_{j=J'}^{J_2} x_{j,p}^{+}(t_k)$$
$$x_p^{+,\text{different}}(t_k) = \frac{1}{J - J_2} \sum_{j=J_2+1}^{J} x_{j,p}^{+}(t_k) \tag{17}$$

Same follows for *minus* neurons. The signal is then averaged across permutations, getting $x^{+,\text{same}}(t_k)$ and $x^{+,\text{different}}(t_k)$ as the signal for *plus* subnetwork and $x^{-,\text{same}}(t_k)$ and $x^{-,\text{different}}(t_k)$ as the signal for the *minus* subnetwork. The significance is evaluated with the permutation test. The test statistic is the sign-specific difference of signals in conditions "same" and "different".

$$\Delta x^{+}(t_k) = x^{+,\text{same}}(t_k) - x^{+,\text{different}}(t_k)$$
$$\Delta x^{-}(t_k) = x^{-,\text{same}}(t_k) - x^{-,\text{different}}(t_k) \tag{18}$$

The null model is computed with the random permutation of class labels 1) when training the classification model and 2) in the reconstruction step. In addition, we use a random assignment to the class of *plus* and *minus* neurons by randomly permuting neural indexes.

The same methods is used to compute the population signal for informative and uninformative neurons, for bursty and non-bursty neurons and in cortical layers.

Correlation function between the population signals

The correlation function between the population signals of *plus* and *minus* subnetworks in trial j and for the permutation instance p is defined as follows:

$$R_{j,p}^{+-}(\tau) = \begin{cases} \sum_{k=0}^{K-\tau-1} x_{j,p}^{+}(t_k) x_{j,p}^{-}(t_k + \tau) & \text{for } \tau \geq 0 \\ R_j^{-+}(-\tau) & \text{for } \tau < 0 \end{cases} \tag{19}$$

with time lag $\tau = 1, 2, ..., 2K - 1$. The correlation function is normalized with autocorrelation functions at zero time lag,

$$\tilde{R}_{j,p}^{+-}(\tau) = \frac{R_{j,p}^{+-}(\tau)}{R_{j,p}^{--}(0) R_{j,p}^{++}(0)} \tag{20}$$

where R^{++} (R^{--}) is the autocorrelation function for *plus* (*minus*) neurons.

$$R_{j,p}^{++}(\tau) = \begin{cases} \sum_{k=0}^{K-\tau-1} x_{j,p}^{+}(t_k) x_{j,p}^{+}(t_k + \tau), & \text{for } \tau \geq 0 \\ R_{j,p}^{++}(-\tau) & \text{for } \tau < 0 \end{cases} \tag{21}$$

The correlation function is computed in single trials and then averaged across trials and across permutations. Since there was no difference in the correlation across conditions, we used all trials from the reconstruction set (conditions CNM and INM, trials with index $j = J_1 + 1, ..., J$).

$$\tilde{R}^{+-}(\tau) = \frac{1}{N_{perm}(J - J_1)} \sum_{p=1}^{N_{perm}} \sum_{j=J_1+1}^{J} \tilde{R}_{j,p}^{+-}(\tau) \tag{22}$$

The significance of the correlation function is estimated with the permutation test. We compute the population signal with random weights and random class label of spike trains. In addition, we use a random assignment to the group of *plus* and *minus* neurons as we compute the correlation function. The same method is used to compute the correlation function between informative and uninformative subnetworks, and between bursty and non-bursty subnetworks.

Correlation function of the population signals in cortical layers

Similarly, we compute the cross-correlation of population signals between pairs of cortical layers. The correlation function is computed for the population signals from two cortical layers,

$$R_{j,p}^{c_1c_2}(\tau) = \begin{cases} \sum_{k=0}^{K-\tau-1} x_{j,p}^{c_1}(t_k) x_{j,p}^{c_2}(t_k+\tau), & \text{for } \tau \geq 0 \\ R_{j,p}^{c_2c_1}(-\tau), & \text{for } \tau < 0 \end{cases} \tag{23}$$

with $(c_1, c_2) \in \{(SG, G), (SG, IG), (G, IG)\}$. The rest of the procedure is the same as for *plus* and *minus* neurons. The significance of results is evaluated with the permutation test, where signals are computed with random weights and random class of the spike train. During the computation of the correlation function, the null model uses random assignment to one of the three cortical layers.

Population vector in the context of *stimulus+choice* and in the context of *choice*

Similarity of population vectors We compare the population vectors in the context of *stimulus+choice* and in the context of *choice*. We train the linear SVM and compute the population vector in each of the two contexts, getting two vectors, $\tilde{\mathbf{w}}^{S+C}$ and $\tilde{\mathbf{w}}^{C}$, in each recording session. Decoding model in the context of *stimulus+choice* utilizes trials from conditions CM and CNM, while decoding model in the context of *choice* utilizes trials from conditions CNM and INM. The number of trials is imbalanced across conditions (namely, there are less trials in condition INM compared to CM), and such imbalance can affect the population vector. We balance the number of trials with the bootstrap method. In each recording session, we find the number of trials of the condition with most trials, and randomly sample, with repetition, the same number of trials from the two other distributions. All reported results are averaged across bootstraps.

We measure the similarity of the two population vectors by computing the angle between them,

$$\alpha = \arccos\left(\tilde{\mathbf{w}}^{S+C} \cdot \tilde{\mathbf{w}}^{C}\right) \tag{24}$$

where $(\cdot)$ is the dot product between the two vectors. Notice that, since vectors are normalized, we have that $||\tilde{\mathbf{w}}^{S+C}|| = ||\tilde{\mathbf{w}}^{C}|| = 1$. If the vectors $\tilde{\mathbf{w}}^{S+C}$ and $\tilde{\mathbf{w}}^{C}$ are similar, they point in similar direction, and the angle between them is small. If, conversely, the two vectors are pointing in random directions between 0 and π, the angle between them is, on average, orthogonal (the average is across bootstrapped samples). The significance of the angle is evaluated with the permutation test, using the test statistics of the angle, averaged across recording sessions. To construct the null model, we draw random vectors from the uniform distribution that have the same range as the true population vectors, and compute the angle between the two random vectors, $\alpha_p = \arccos\left(\tilde{\mathbf{w}}_p^{S+C} \cdot \tilde{\mathbf{w}}_p^{C}\right)$, $p = 1, ..., N_{perm}$. The p-value is computed by ranking the angle of the true model among the distribution of angles of the model with random weights. The test is significant if $p < {}^{0.05}/_{N_{test}}$, where the division with the number of tests implements the Bonferroni correction for multiple testing.

Distance of decoding weights While the similarity of population vectors is a population-wise measure, we are also interested in understanding how different are weights of single neurons in the context of *stimulus+choice* and in the context of *choice*. For each single neuron, we compute the distance between its decoding weight in

the two contexts.

$$d_n^{S+C,C} = |\tilde{w}_n^{S+C} - \tilde{w}_n^{C}| \tag{25}$$

Results are collected across recording sessions. We then split neurons in groups according to a specific criterion (sign of the weight, informativeness, burstiness, layers) and test the difference between groups with a two-tailed t-test.

Results

Learning representation on *stimulus+choice* transfers to the representation of *choice*.

Two adult male macaques were trained on a delayed match-to-sample visual task. The subject visualized the target and the test stimuli, with a delay period in between (fig. 1A). The target and test stimulus were either identical (condition match) or not (condition non-match), with the only possible difference being the change in the orientation of the test stimulus. The subject had to report its decision about the stimulus class ("same" or "different") and was rewarded for correct behavior (see methods). The stimuli were complex naturalistic images in black and white, depicting an outdoor scene, and their identity changed from one trial to another. We recorded the multiunit signal in V1 with a linear array, capturing the activity across the cortical depth. After spike sorting, we obtained on the order of 10 units in each recording session, and the total of 160 units in 20 recording sessions. For example spike rasters, see fig. 1B. For further specifications on the task and on recordings, see methods.

Single neuron statistics and standard decoding methods on parallel spike trains from populations of simultaneously recorded neurons do not allow to predict animal's choice. Firing rates of single neurons vary strongly across units, but are very similar for the choice "same" and "different" (fig. 1C). Coefficient of variation (CV_2, see methods) of single neurons varies around 1, which is the CV_2 of the Poisson process, and is strongly positively correlated across conditions (fig. 1D). Decoding from parallel spike counts with linear Support Vector Machine gives chance level performance (balanced accuracy of 0.505, 100 cross-validations). Similarly, a temporally resolved method, the population peri-stimulus time histogram, gives highly overlapping signals for choices "same" and "different" (fig. 2A). In the previous work, we suggested a decoding method where we learn structural features of the activity of neural populations on parallel spike counts, and then use these features to weight parallel spike trains and compute temporally resolved population signal in single trials ([14]). While the method has been successful in predicting the correct choice behavior (i.e., the *stimulus + choice* signal, see fig. 2 in [14]), it only gives a poor prediction of the choice in absence of the information on the stimuli (i.e., the *choice* signal, fig. S1.1).

The biological interpretation of a decoding weight is to represent the strength of the synaptic connection between the presynaptic and the post-synaptic neuron. In this sense, structural features of the population activity (or, decoding weights) can be interpreted as synaptic weights that are stronger (positive weights) or weaker (negative weights) than the baseline synaptic weight ([14]). This can be achieved through biologically plausible learning, with reward as the teaching signal. Since the subject is only rewarded in correct trials, we reason that a simple way of learning decoding weights in V1 could be limited to correct trials. The representation, however, must also take place in incorrect trials. We therefore learn decoding weights in correct trials, where there is information about both the stimuli and the choice (the *stimulus + choice* signal), and then compute the read-out from parallel spike trains from trials that only differ in the choice, but not in the stimulus class (the *choice* signal, see table 1, see methods). We refer to the decoded signal as the population signal, that represents the synaptic current of a hypothetical read-out neuron. If not stated otherwise, we always use the time window [0,400] ms after the onset of the test stimulus.

Surprisingly, we find that with such learning, the choice signal for "same" vs. "different" can be discriminated during the presentation of the test stimulus (fig. 2B), with the probability better than chance (fig. 2B-C). The

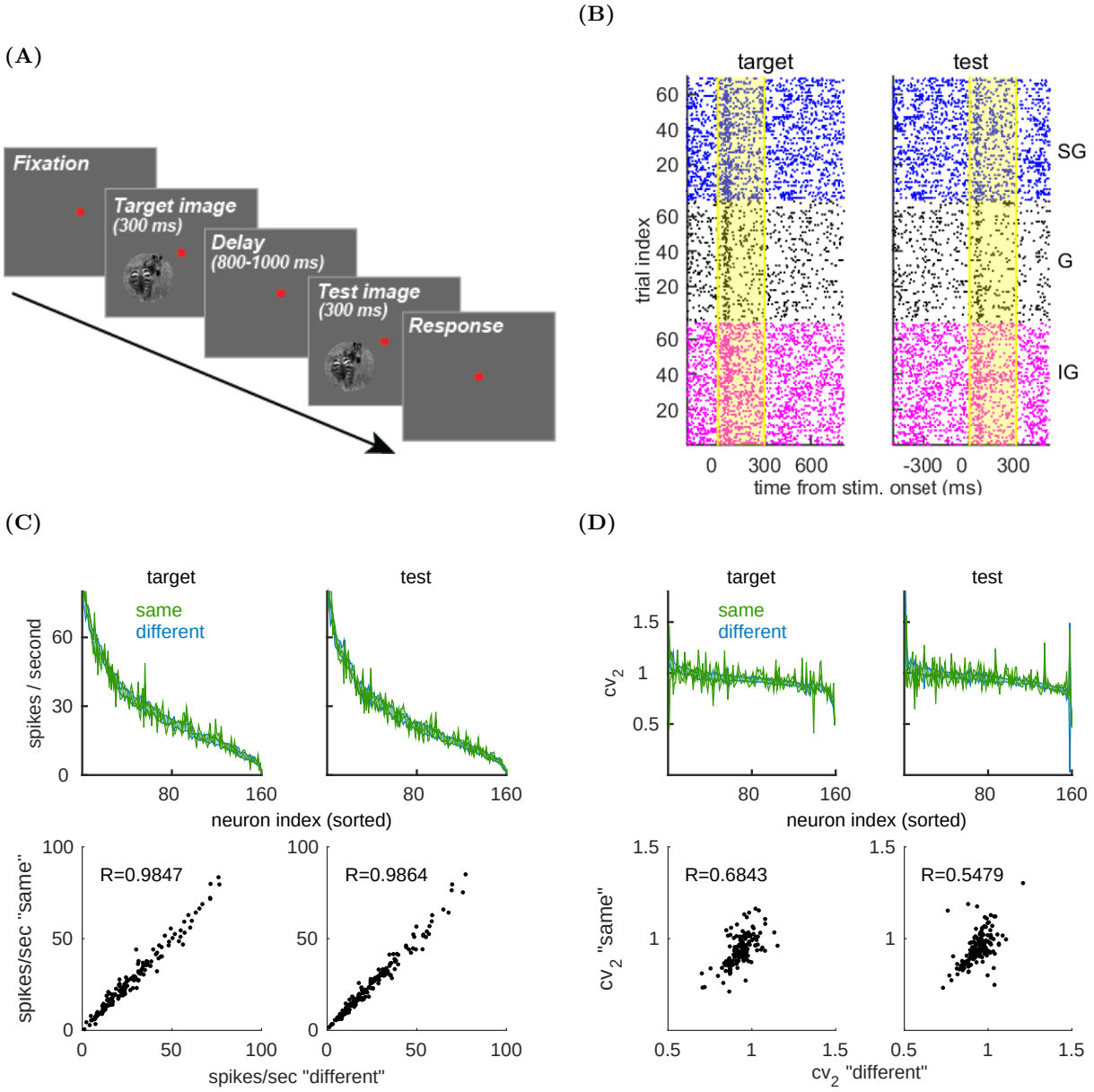

Figure 1. Experimental paradigm and single neuron statistics. **(A)** Experimental paradigm. The subject visualizes the target and the test stimulus, with a delay period in between. **(B)** Spike raster of three representative single neurons for the behavioral choice "different". We show a spike raster of three neurons, one in superficial layer (SG, blue), one in the middle layer (G, black) and one in the deep layer (IG, magenta). **(C)** Top: Firing rate (mean ± SEM) for single neurons during target (left) and test (right), corresponding to the choice "same" (green) and "different" (blue). Results are sorted w.r.t. the firing rate for the choice "different". The SEM is for the variability across trials and results are computed in the time window [0,400] ms w.r.t. the stimulus onset. Bottom: Scatter plot of the mean firing rate for the choice "different" (x-axis) and for the choice "same" (y-axis). "R" is the Pearson correlation coefficient. **(D)** Same as in **C**, but showing the coefficient of variation (CV_2).

population signal depends only weakly on the time scale of the convolution, with longer time scales resulting in a smoother signal (fig. 2D). Such decoding is also robust to the length of the time window that we use for decoding (fig. S1.2). We therefore conclude that learning in the context of *stimulus + choice* successfully transfers to the representation of *choice* on the hold-out test set. During the target time window, the information that is necessary for discrimination is not yet available. Accordingly, we find that the population signal stays close to zero and, as expected, prediction of the choice behavior is not possible (fig. S1.3B-C). Interestingly, the read-out of the choice during test switches at about 140 ms after the stimulus onset (fig. 2C, red arrow). We notice that this corresponds to the moment where the activity qualitatively changes from transient response after the stimulus onset to tonic firing thereafter (fig. 2A, red arrow).

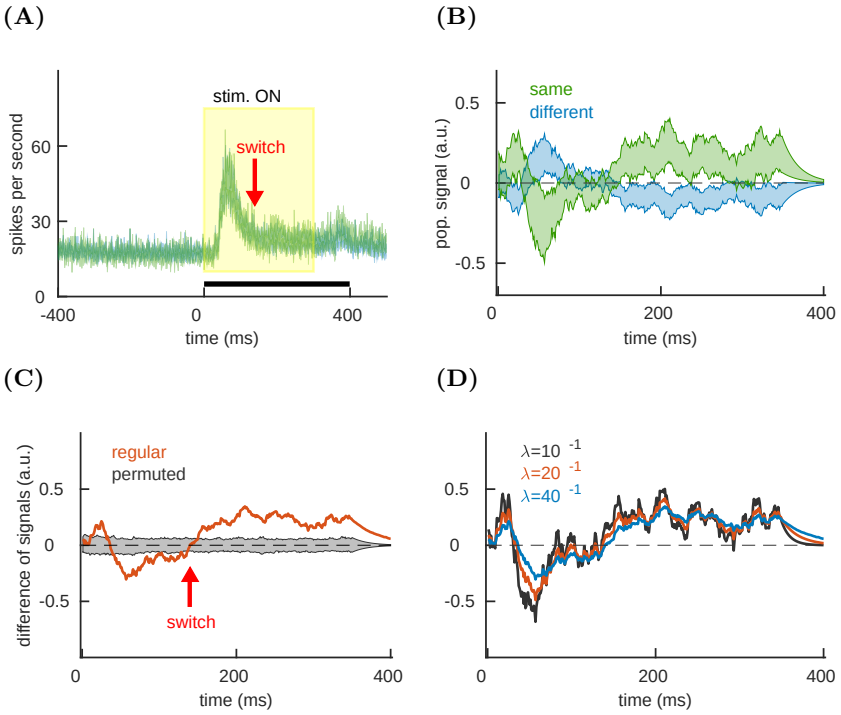

Figure 2. Behavioral choice can be discriminated with the transfer of learning. **(A)** Population PSTHs in for the choice "same" (green) and "different" (blue). We show the mean ± SEM for the variability across recording sessions. The yellow window indicates the presence of the stimulus and the black bar indicates the time window, used for analysis. The red arrow marks the time of the zero crossing of the population signal. **(B)** Population signal for the choice "same" (green) and "different" (blue). We show the mean ± SEM for the variability across sessions. Parameter: $\lambda^{-1} = 20$. **(C)** Session-averaged difference of population signals for the regular model (red) and the model with permuted class labels (gray). Parameters: $\lambda^{-1} = 20\ ms$, $N_{perm} = 1000$. **(D)** Session-averaged difference of the population signals for three values of the time constant of the convolution.

Discrimination relies on the sign on of decoding weight as well as on the timing of spikes.

The read-out model that we propose here is a linear weighted sum of spikes, going through a convolution (see methods). In spite of its simplicity, this linear decoder has several components that may or may not be necessarily required for discrimination. In order to test the necessity of each component for discrimination, we remove the information of a specific component and test the effect of such a perturbation on discrimination. First, we randomize the sign of decoding weights (see methods). Randomizing the sign clearly destroys the discriminatory power of the model (fig. 3A, top left), meaning that the sign of the weight is required for discrimination. Next, we keep the correct sign but randomize the amplitude of the entries of the weight vector. This perturbation does not compromise the discrimination (fig. 3A, top right), implying that the information on the amplitude of weights is not crucial for discrimination. As we permute the neural index across neurons,

we get a noisier read-out, that is, however, still predictive of the choice (fig. 3A, bottom left). In contrast, the discriminatory capacity of the model is entirely lost when we randomly permute the class label ("same" and "different") of parallel spike trains in the reconstruction step (fig. 3A, bottom right). These results demonstrate that the sign of the weight (in the learning step) and the correct class label for spike trains (in the reconstruction step) and necessarily required for discrimination of the choice signal.

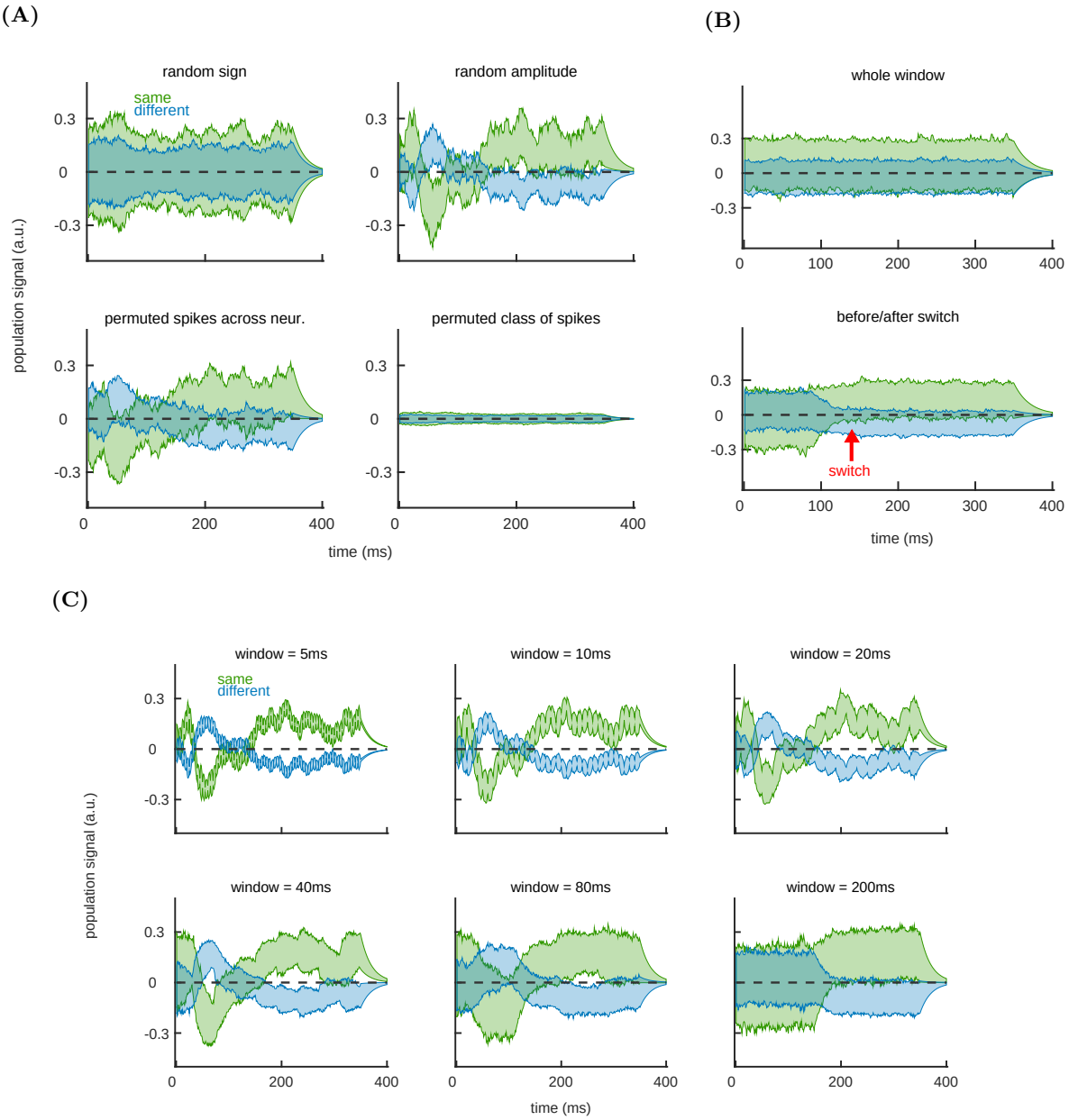

Figure 3. Correct sign of weights and spike timing are necessary for discrimination of the choice. (A) Population signals with random sign of weights (top left), random amplitude of weights (top right), permuted spikes across neurons (bottom left) and permuted class labels of spike trains (bottom right). We plot the entire distribution of population signals in conditions "same" (green) and "different" (blue). **(B)** Permuted spike timing across the entire time window (top) and in two time windows, before and after the switch in representation (bottom). Red arrow marks the timing of the switch (140 ms after the stimulus onset). **(C)** Population signal with temporal jitter, for different length of the jittering window. Parameters for all plots: $\lambda^{-1} = 20\ ms$, $N_{perm} = 1000$.

Next, we test the necessity of correct spike timing for discrimination. As we randomly permute the spike timing across the entire time window of 400 ms, we get highly overlapping signals for conditions "same" and "different" (fig. 3B, top). Interestingly, if the spike timing is permuted in the time window only before and

after the switch in representation (fig. 2A and C, red arrow), the overlap of the population signals is largely diminished (fig. 3B, bottom). This suggests different sources of (the informative part of) the population signal before and after the switch, and implies that the information in firing rate is informative for discrimination if we separate these sources. Finally, we ask how does the temporal jitter in smaller time windows affect the read-out. Surprisingly, the temporal jitter in shorter time windows makes the choice signal noisier, but still allows discrimination (fig. 3C).

The choice signal of neurons with positive and negative weights is negatively correlated.

The sign of the weight is an important characteristics of decoding weights, since neurons with the opposite sign of the weight have the opposite effect on the classification model, pulling the separation boundary in opposite directions. We separate the population with respect to the sign of the weight, getting *plus* and *minus* subnetworks. As we compute the population signal independently with each subnetwork (see methods), we find that the *plus* subnetwork discriminates the choice behavior better than the *minus* subnetwork (fig. 4A). The population signal of the *plus* subnetwork also accounts for the switch in representation at around 140 ms after the stimulus onset (fig. 4A, right). The *minus* subnetwork shows some discriminatory capacity at the very end of the trial (fig. 4A, left), where the response is in anti-phase with respect to the signal of the *plus* subnetwork. In fact, the signal of *minus* neurons increases for the choice "different" while the signal of *plus* neurons decreases for the same choice. Accordingly, the population signal of *plus* and *minus* subnetworks is negatively correlated (see methods for the definition of the correlation function), consistently for different time constants of the convolution (fig. 4B, top, table S2.1). This effect is strongly significant (fig. 4B, bottom).

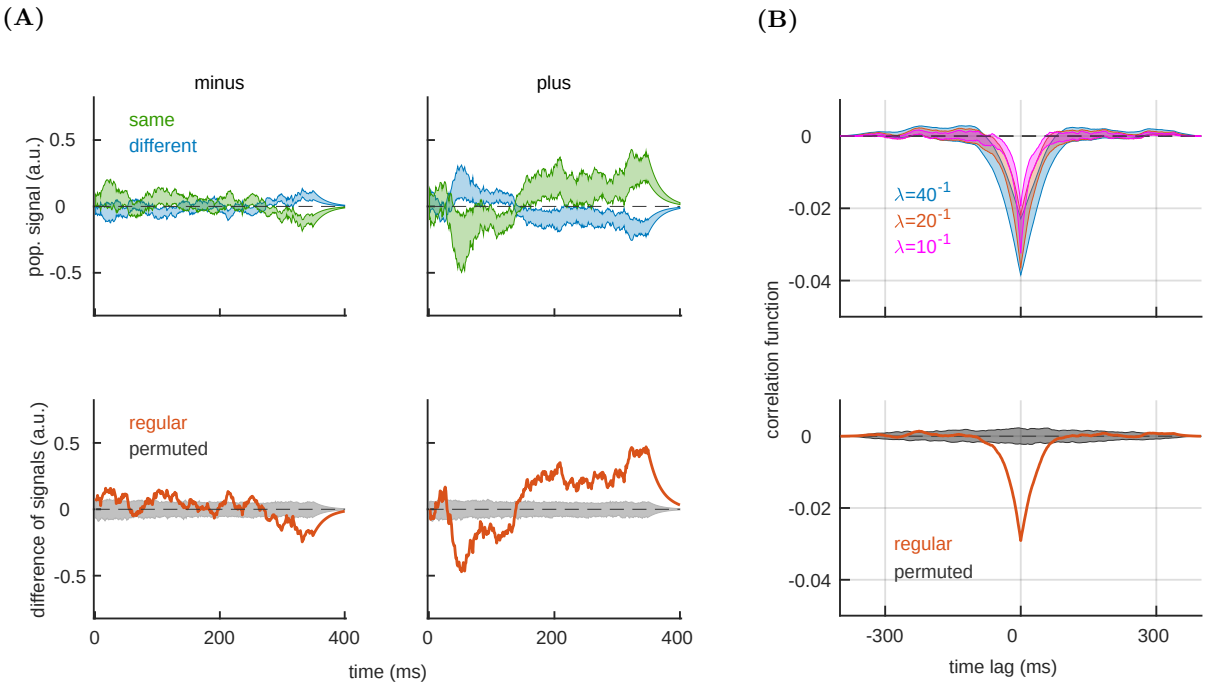

Figure 4. Subpopulations of *plus* and *minus* neurons are negatively correlated. (A) Top: Population signal for *minus* neurons (left) and for *plus* neurons (right). We show the mean ± SEM for the variability across recording sessions. Bottom: Session-averaged difference of signals for the true model (red) and for models with permutation (gray area). **(B)** Top: Correlation function between the population signal of *plus* and *minus* subnetworks for different values of the time constant of the convolution λ. Bottom: Session-averaged correlation function for the regular model (red) and the distribution of results for the model with permutation (gray area). Parameters for all plots, when applicable: $\lambda^{-1} = 20\ ms$, $N_{perm} = 1000$.

Only the activity of a subset of informative neurons is relevant for discrimination.

Besides the sign of the weight, the strength of the weight, $|\tilde{w}_n|$, is an important characteristics of the decoding model, since neurons with strong weight are those that contribute most to discrimination. We separate neurons in informative and uninformative, by ranking the strength of the weight of each neuron among the distribution of weights of models with permuted class labels (see methods). As we compute the population signal for informative and uninformative subnetworks, not surprisingly, informative subnetwork is the one that carries the choice related signal, while the network of uninformative neurons performs at chance (fig. 5A). The correlation function between signals of informative and uninformative neurons shows that the two signals are positively correlated (fig. 5B, table S2.2).

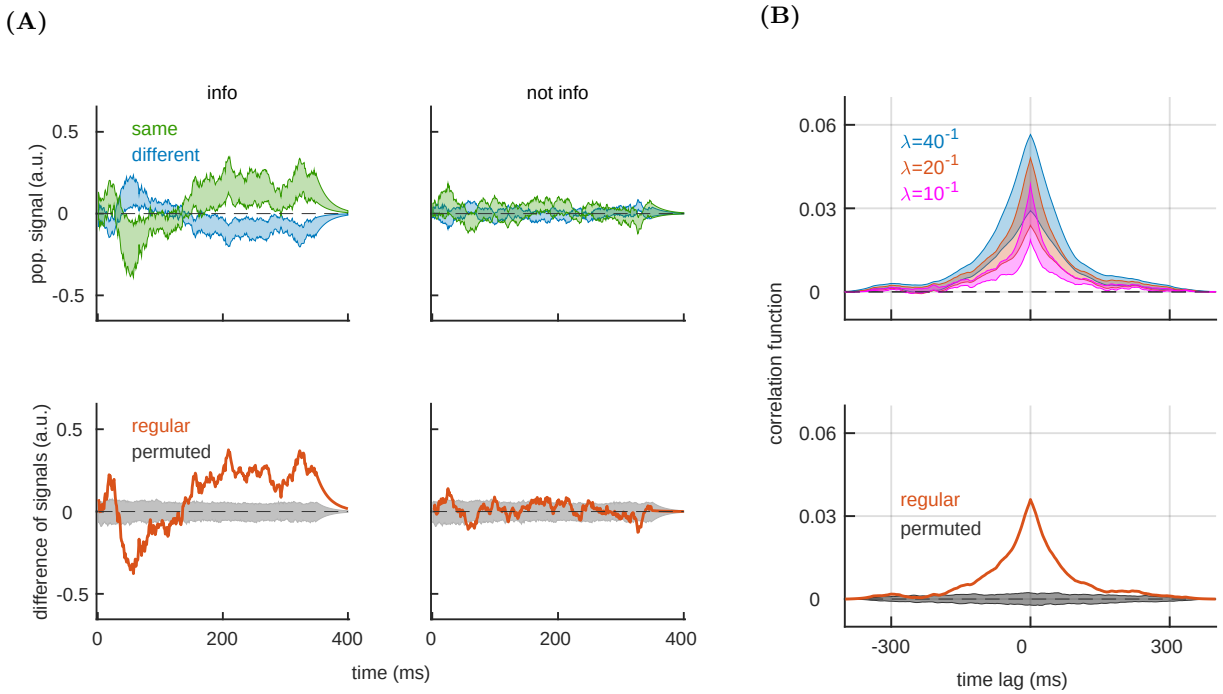

Figure 5. Discrimination relies on a subset of informative neurons. **(A)** Top: Population signal of subnetworks of informative (left) and uninformative neurons (right), showing the mean ± SEM for the variability across recording sessions. Bottom: Session-averaged difference of signals for the true model (red) and for models with permutation (gray area). **(B)** Top: Correlation function between the population signal of informative and uninformative neurons for different values of the time constant of the convolution. We show the mean ± SEM for the variability across recording sessions. Bottom: Session-averaged correlation function for the regular model (red) and the distribution of results for the model with permutation (gray area). Parameters for all plots, when applicable: $\lambda^{-1} = 20\ ms$, $N_{perm} = 1000$.

Activity of bursty neurons is particularly informative about the choice.

Next, we divide neurons into bursty and non-bursty, based on characteristics of the power spectrum of spike trains (see methods). The power spectrum of bursty neuron is characterized by a reduced power for middle range frequencies ([6], see fig. 6A, middle right and right in the top row). We capture this effect by computing the normalized area under the power spectrum. We define bursty neurons as those that have the area under the power spectrum smaller than the lower bound of the same result for models with the permuted spike timing (random permutation of the time index, neurons with red asterisk on fig. 6B). As we compute the population signal for bursty and non-bursty subnetworks, we find that the bursty subnetwork is better at predicting the choice signal than the non-bursty subnetwork (fig. 6C). The correlation function between the population signals of the two subnetworks shows that the spiking activity in bursty and non-bursty subnetwork is positively correlated (fig. 6D, table S2.3).

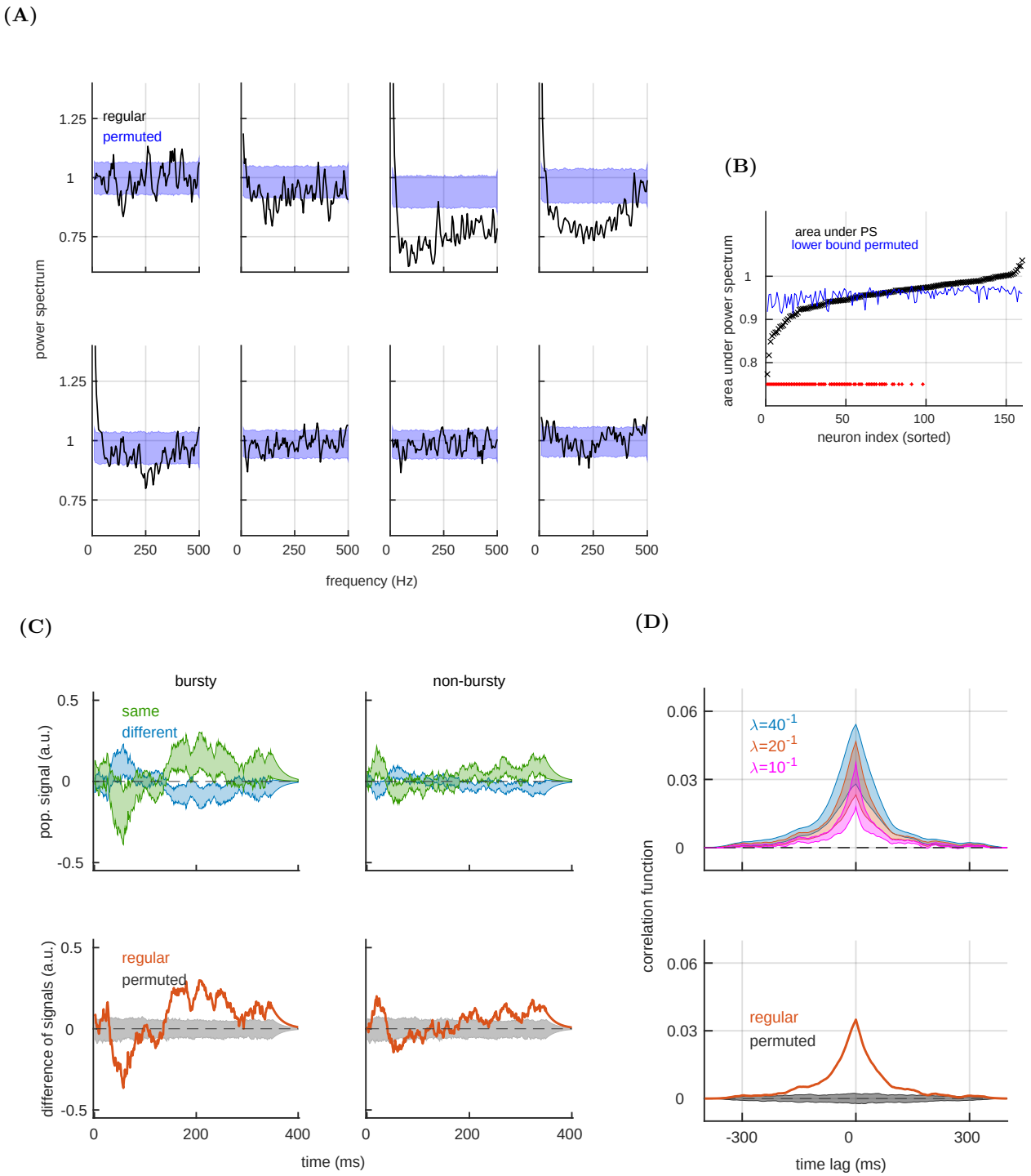

Figure 6. Bursty neurons discriminate better and earlier in the trial. (A) Power spectrum for 8 neurons in an example recording session. We show the power spectrum for the true model (black) and the distribution of results for models with permuted spike timing (blue). Bursty neurons are those with reduced power for lower and medium range frequencies (top row, middle left and left). **(B)** Area under the power spectrum for all neurons, sorted from lowest to highest value (black), and the lower bound of the area under the power spectrum for neurons with permuted time index (blue). Red asterisk marks significance (permutation test, $\alpha = 0.05$). **(C)** Top: Population signal for bursty (left) and non-bursty subnetwork (right) in recording sessions. Bottom: Session-averaged difference of signals for the true model (red) and for models with permutation (gray area). **(D)** Top: Correlation function between the population signal of bursty and non-bursty subnetworks for different values of the time constant of the convolution, λ. We show the mean $\pm$ SEM for the variability across recording sessions. Bottom: Session-averaged correlation function for the regular model (red) and the distribution of results for the model with permutation (gray area). Parameters for all plots, when applicable: $\lambda^{-1} = 20\ ms$, $N_{perm} = 1000$.

The choice signal differs across cortical layers.

Cortical layers in V1 have been shown to importantly differ in their correlation structure [10], which might have direct implications on the population code. Using a method based on current source density (see methods, [14]), we distinguish the superficial layer (supragranular, SG), the middle layer (granular, G) and the deep layer (infragranular, IG). In the previous work, we showed that the quality of discrimination of the *stimulus + choice* variable differs across layers, with superficial layers showing the strongest discriminatory capacity. We ask whether the same effect is observed with transfer of learning to the *choice* variable. We compute the population signal in each layer (see methods) to found out that the superficial layer is the most discriminative of the *choice* signal (fig. 7A). The superficial layer also reflects the switch of the representation at around 140 ms after the stimulus onset, seen in the overall response (compare fig. 7A, top right, with fig. 2C). Interestingly, at the beginning of the trial, the representation in the superficial layer seems to be in disagreement with the representation in the middle and deep layers. Nevertheless, the population signals is weakly positively correlated across all pairs of laminar circuits (fig. 7B), with the positive correlation being consistent for different choice of the time constant of the convolution (fig. 7B, left). The strongest correlation is observed between the middle and the superficial layer (for $\lambda = 20\ ms^{-1}$, peak=[0.014, 0.012, 0.010] for SG/G, SG/IG and G/IG, respectively, see table S2.4). This is consistent with the hypothesis that the visual signal enters the middle cortical layer and is then projected to the superficial layer, where top-down and bottom-up signals are compared. In addition, there is excess of *plus* neurons and bursty neurons in the superficial layer (fig. 8A), which might explain why the superficial layer has stronger discriminatory capacity compared to other layers.

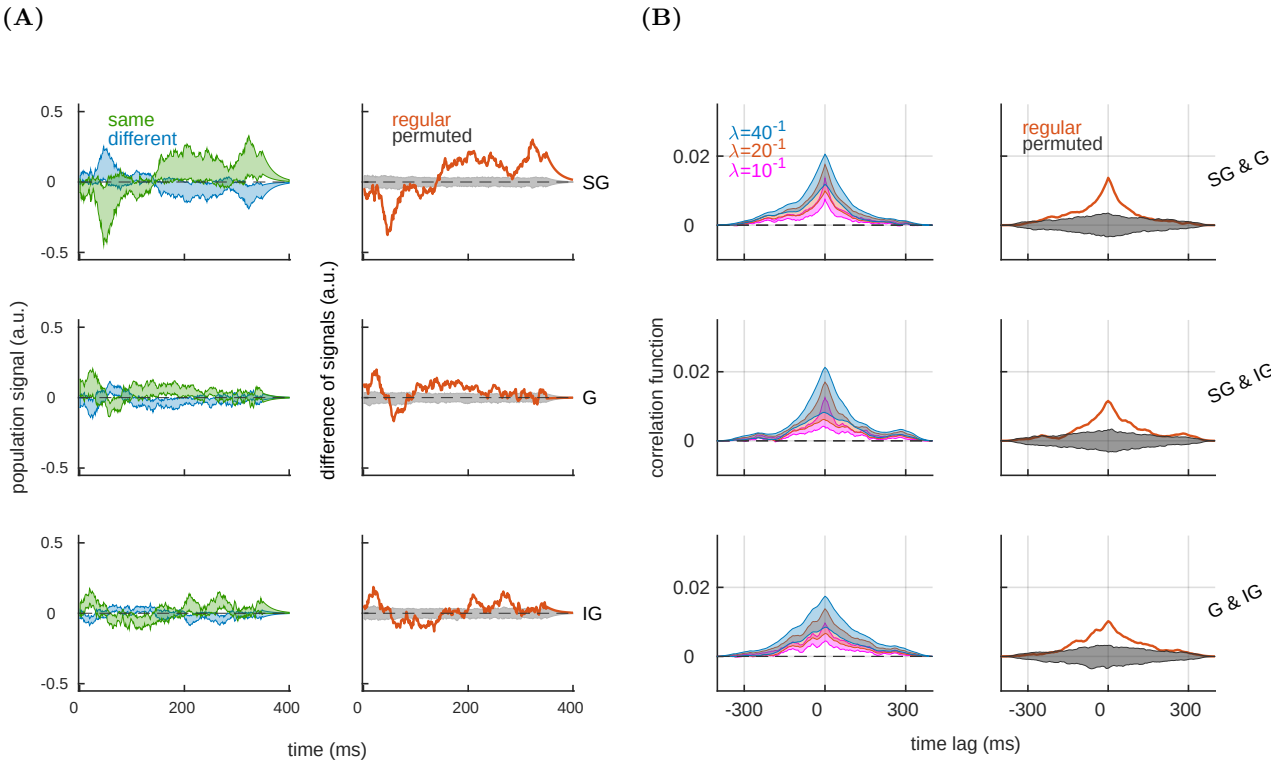

Figure 7. Activity in the superficial layer of the cortex is the most discriminative of the choice behavior. (A) Left: Population signal in the SG (top), G (middle) and IG layer (bottom). We show the mean ± SEM across recording sessions. Right: Session-averaged difference of population signals for the true model (red) and the model with permuted class labels (gray). Parameters: $\lambda^{-1} = 20\ ms$, $N_{perm} = 1000$. **(B)** Left: Correlation function of the population signal across pairs of layers. We show the mean ± SEM across recording sessions and plot results for three values of the time constant of the convolution. Right: Session-averaged correlation function for the regular model (red) and for models with permutation (gray area). We use the time constant $\lambda^{-1} = 20\ ms$ and $N_{perm} = 1000$ random permutations.

Informative neurons are more sensitive to the informational context than uninformative neurons.

In the last section, we compare population vectors in the context of *stimulus* + *choice*, $\tilde{\mathbf{w}}^{S+C}$, with population vectors in the context of *choice*, $\tilde{\mathbf{w}}^{C}$. While only the former might be relevant for decoding of the population signal, comparison of population vectors across the two informational contexts allows to better understand how neuronal responses change from one context to the other (table 1). Population vectors are computed in each recording session (see methods). First, we estimate the similarity of population responses across the two classification problems with the angle between the two population vectors (fig. 8B, see methods). If population vectors are independent, they point in random direction between 0 and π, and their average is an orthogonal angle, $\alpha^{\text{independent}} = \pi/2$. Deviation from orthogonality indicates the similarity of weight vectors $\tilde{\mathbf{w}}^{S+C}$ and $\tilde{\mathbf{w}}^{C}$. We find that, on average, the angle between the two weight vectors weakly but significantly deviates from the orthogonal angle (fig. 8B). This means that the two weight vectors are not independent, but weakly similar. This result is robust to the length of the time window ($\alpha = [80.2,\ 80.8,\ 80.0]$ degrees for the length of the time window $K = [300,\ 400,\ 500]$, respectively) and is significant in all cases ($p < 0.001$, in all cases, permutation test with 1000 permutations).

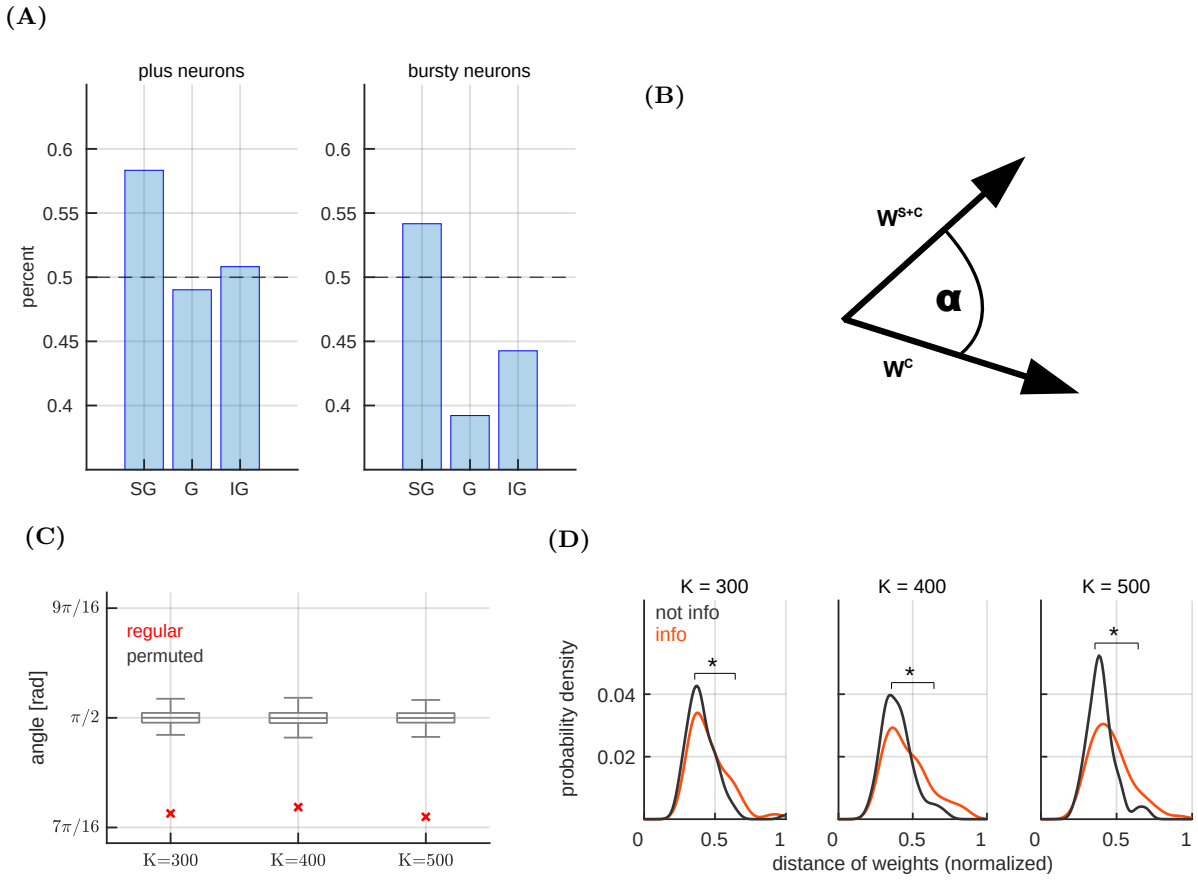

Figure 8. The difference between the representation of *stimulus+choice* and *choice* is stronger in informative than in uninformative neurons. **(A)** The proportion of *plus* neurons (left) and of bursty neurons (right) in cortical layers. The proportion is computed in each layer individually. As an example, the proportion of 0.4 of bursty neurons in the G layer means that, from neurons in the G layer, 40 % of neurons are bursty and 60 % of neurons are non-bursty. **(B)** Schema of the angle between the population vectors in the context of *stimulus* + *choice* and *choice*. **(C)** Average angle (red) and the distribution of results for models with random weights (gray box-plot), for the length of the time window $K = 300$ (left), $K = 400$ (middle) and $K = 500$ (right). Parameter: $N_{perm} = 1000$. **(D)** Distance of weights between the two classification problems for informative (orange) and uninformative (black) neurons. We show results for the length of the time window $K = 300$ (left), $K = 400$ (middle) and $K = 500$ (right). The asterisk marks significant two-tailed t-test.

Finally, we compare decoding weights of single neurons between the contexts of *stimulus+choice* and *choice* by measuring the distance of the decoding weight in two classification problems (see methods). The distance of decoding weight is measured for each neuron, and collected across recording sessions. We then divide neurons in informative and uninformative (see methods) and test whether the two groups differ in their sensitivity to the informational context. We find that informative neurons are more sensitive to the informational context than uninformative neurons ($p = 0.0184$ for K=300, $p = 0.0017$ for K=400, $p = 0.0007$ for K=500, two-tailed t-test; fig. 8D). As neurons are split according to the sign of the weight and with respect to the burstiness criterion, there is no significant difference across groups (fig. S3.1A-B). Same is true for the division in cortical layers, that also do not show any significant difference in their sensitivity to the informational context (fig. S3.1C).

Discussion

In the present work, we decoded binary choice signal from parallel spike trains in the primary visual cortex of the macaque. We trained the read-out in the context of *stimulus + choice* and validated the model in the context of *choice*, hypothesizing that successful discrimination in this setting will show the intersection between the two informational contexts. Results showed that the behavioral choice can be decoded from V1 in the context of *choice*. This strongly suggests that, in the present experimental setting, the representation of the stimulus class has a substantial non-zero intersection with the representation of the behavioral choice, in the sense of [26].

The origin of the choice signal

The origin of the choice signal that we decode from present dataset is a matter of discussion. It has been proposed that the choice-related signal (that cannot utilize the difference in stimuli) might originate from feed-forward projections [22], feed-back projections [25], or be due to local network dynamics (see [41, 42] for modeling studies). In the present analysis, a piece of evidence that allows to discuss about the sources of the choice signal is the switch in representation at about 140 ms after the stimulus onset, which coincides with the qualitative change of the population response from the strong transient at the stimulus onset to tonic firing thereafter. At the stimulus onset, the neural activity is expected to be mainly driven by the feed-forward input, while feed-back and local projections are expected to have have stronger effect, respectively to the feed-forward ones, during the second half of the trial [4]. If the representation in the context of *choice* would be only due to the feed-forward drive throughout the trial, we would expect that the representation remains consistent throughout the trial. Also, we would expect it to be consistent between the middle and the superficial layer, which, at least at the beginning of the trial, is not the case. We propose that the switch in representation is caused by the feed-back and local projections taking over the feed-forward drive, where the switch would be due to incongruent signals between the bottom-up and the top-down pathways.

We suggested incongruent top-down and bottom-up signals in the context of *choice*. It follows that in the context of *stimulus + choice*, when the behavior of the animal is correct for both choices, the feed-forward and the feed-back drive are expected to be congruent in their signaling, giving consistent representation of the choice throughout the trial and across layers. In the previous work, we have shown that this is indeed the case ([14]). Another result that suggests the existence of both bottom-up and top-down sources is related to the spike timing. In the present work, we have shown that jittering the spike timing in small windows only makes the population signal noisier, but does not entirely destroy the discriminability of the choices. Permuting the spike timing across the entire window has destroyed the discriminability of the choice signal, while limiting the window of permutation to the time before and after the switch in representation, the discriminability is partially preserved. This supports the hypothesis of different origins of the signal before and after the switch. If the representation of the choice in V1 partially relies on feedback projections, the timing of the feedback might be crucial for perception, as recently demonstrated with the feedback projection from V5 to V1 on visual perception [37].

The third piece of evidence in favor of the importance of lateral and feed-back connections is the condition-specific divergence of population signals for neurons with positive and negative weights towards the end of the trial. This effect can be explained by a top-down input that selectively drives neurons with positive weight for the decision "same" and neurons with negative weight for the decision "different". It has been shown that such a a context-dependent computation is feasible in a recurrent network [19] and it could be implemented, in the prefrontal cortex. As a first alternative, a top-down signal could drive bursty neurons in the superficial layer through dendritic feedback inhibition [21]. As a second alternative, coding pools of *plus* and *minus* neurons could be formed locally, and arise due to the structure of the lateral connectivity of the local network.

Relation to other approaches

Decoding in the context of *choice* is similar to computing the choice probability ([35, 3, 34, 17, 29, 24, 25]), where prediction of the behavioral choice is also measured in the absence of the information about the stimuli. In the MT area of the macaque, previous studies reported the independence of the choice probability on the coherence of the stimulus [3, 40]. Our results, in contrast, demonstrate that V1 neurons are sensitive to the presence of the information about the stimuli, in particular neurons that have a strong decoding weight (informative neurons). We showed that population vectors in the context of $stimulus + choice$ and in the context of *choice* are weakly similar, but are far from being entirely aligned. The discrepancy of our results with results of studies in the MT cortex might be due to the fact that MT area is higher in the visual hierarchy than V1, and supposedly "closer" to the representation of the behavioral output. Moreover, studies [3, 40] decode from single neurons while we decode from neural populations, using a different decoding method.

Representation of the behavioral choice in absence of sensory evidence has been extensively studied since around three decades. Studies on choice probability have primarily addressed the time-averaged activity of single neurons and their relation to the behavioral choice of the animal, and single neurons have been shown to correlate with the choice behavior in multiple brain areas [35, 3, 17, 40, 29]. However, even if it is possible to find neurons whose activity predicts the behavior with better-than-chance probability, it is not clear how would the brain isolate the activity of a single neuron from a vast number of cells that activate in parallel. Recent years have seen the revival and the development of the idea that behaviorally relevant variables are encoded by collective dynamics of neural ensembles (see, e.g., [9, 30, 2, 22, 13]). As shown recently in [32, 33], a good description of the spiking activity of a nearly complete network of retinal ganglion cells has to take into account single neuron firing rates, the strength of pairwise interactions and a global modulation of the network activity. Population codes seem to be appropriate to study neural responses in the auditory [11, 18], motor [5], prefrontal [28], and visual cortex [22]. When analyzing the data, the multivariate approaches, that take into account interactions between neurons, seem to be more appropriate for the analysis of parallel activity than univariate approaches [30, 32].

An impressive statistical description of the activity of neural ensembles in the retina has been recently achieved by maximum entropy models [33, 32]. While such models give a statistical description of important dynamical properties of neural responses, they are not directly informative about the neural function, and about the relation between neural activity and behavior. Studying the relation between the neural activity and behavior is precisely the purpose of decoding models, and in this respect, decoding models complement statistical models. A widespread family of decoding models are Models of Neuronal Stimulus-Response Function, that relate the activity of single neurons to selected features of the stimulus (see [20, 12] for a review). While such mapping is informative about the coding function of single neurons, it gives only an incomplete (and potentially biased) description of coding mechanisms in the brain. A particular difficulty for the interpretation of the mapping between features of sensory stimuli and neural responses might arise due to correlations between features. If neurons have mixed selectivity and respond to a mix of sensory features [28], probing neuron's response with only one of many correlated features will elicit an incomplete picture about the neuron's response function. If we assume that neurons are optimized to process natural stimuli, where sensory features tend to be correlated [31], mixed selectivity is expected not only in high-level cognitive areas but throughout the cortex.

With this in mind, it might be more straightforward to relate the activity of neural ensembles to natural stimuli and to variables that describe animal's behavior. Nevertheless, the artificial stimuli can be better controlled and studying neural responses to such stimuli is therefore complementary to studying responses to naturalistic stimuli.

Neurons behind sensory receptors only receive activation of other neurons and are not in direct contact with the stimuli [32]. In this sense, the mapping between the neural activity in a sensory area and the behavioral output does not map a direct chain of events, but two distant elements of that chain, jumping across multiple elements in between. The difficulty of studying causal relations is a long-standing challenge in neuroscience, arising from the difficulty of isolating a specific part of the brain without perturbing its natural working regime. In practice, this difficulty is reflected in technical limitations of experimental neuroscience (but see [16, 1, 37]) and in the widespread use of the correlation measure for the analysis (but see [38, 39]). An even more solid ground for the interpretation of parallel spike trains with respect to behavioral variables would therefore be given by probing causal relations between the activity of neural populations across different stages of processing, still a major challenge for future research.

Financial disclosure

This work was supported by Deutsche Forschungsgemeinschaft, grants GRK 1589/2 and SFB 1315.

References

1. Andrei, A. R., Pojoga, S., Janz, R., & Dragoi, V. (2019). Integration of cortical population signals for visual perception. Nature communications, 10(1), 1-13.

2. Boerlin, M., Machens, C. K., & Denève, S. (2013). Predictive coding of dynamical variables in balanced spiking networks. PLoS computational biology, 9(11), e1003258.

3. Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S., & Movshon, J. A. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. Visual neuroscience, 13(1), 87-100.

4. Callaway, E. M. (2004). Feedforward, feedback and inhibitory connections in primate visual cortex. Neural Networks, 17(5-6), 625-632.

5. Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., & Shenoy, K. V. (2012). Neural population dynamics during reaching. Nature, 487(7405), 51.

6. Compte, A., Constantinidis, C., Tegnr, J., Raghavachari, S., Chafee, M. V., Goldman-Rakic, P. S., & Wang, X. J. (2003). Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. Journal of neurophysiology, 90(5), 3441-3454.

7. Destexhe, A., Rudolph, M., & Par, D. (2003). The high-conductance state of neocortical neurons in vivo. Nature reviews neuroscience, 4(9), 739.

8. Elsayed, G. F., & Cunningham, J. P. (2017). Structure in neural population recordings: an expected byproduct of simpler phenomena?. Nature neuroscience, 20(9), 1310.

9. Gao, P., & Ganguli, S. (2015). On simplicity and complexity in the brave new world of large-scale neuroscience. Current opinion in neurobiology, 32, 148-155.

10. Hansen, B. J., Chelaru, M. I., & Dragoi, V. (2012). Correlated variability in laminar cortical circuits. Neuron, 76(3), 590-602.

11. Harris, K. D., Bartho, P., Chadderton, P., Curto, C., de la Rocha, J., Hollender, L., ... & Sakata, S. (2011). How do neurons work together? Lessons from auditory cortex. Hearing research, 271(1-2), 37-53.

12. Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and decoding models in cognitive electrophysiology. Frontiers in systems neuroscience, 11, 61.

13. Koren, V., & Denève, S. (2017). Computational Account of Spontaneous Activity as a Signature of Predictive Coding. PLoS computational biology, 13(1), e1005355.

14. Koren V, Andrei AR, Hu M, Dragoi V, Obermayer K (2019) Reading-out task variables as a low-dimensional reconstruction of neural spike trains in single trials. PLoS ONE 14(10): e0222649. https://doi.org/10.1371/journal.pone.0222649

15. Veronika Koren, Ariana R. Andrei, Ming Hu, Valentin Dragoi, Klaus Obermayer bioRxiv 645135; doi: https://doi.org/10.1101/645135, *preprint*

16. Ledochowitsch, P., Yazdan-Shahmorad, A., Bouchard, K. E., Diaz-Botia, C., Hanson, T. L., He, J. W., ... & Schreiner, C. E. (2015). Strategies for optical control and simultaneous electrical readout of extended cortical circuits. Journal of neuroscience methods, 256, 220-231.

17. Logothetis, N. K., & Schall, J. D. (1989). Neuronal correlates of subjective visual perception. Science, 245(4919), 761-763.

18. Luczak, A., Barth, P., & Harris, K. D. (2009). Spontaneous events outline the realm of possible sensory responses in neocortical populations. Neuron, 62(3), 413-425.

19. Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. nature, 503(7474), 78.

20. Meyer, A. F., Williamson, R. S., Linden, J. F., & Sahani, M. (2017). Models of neuronal stimulus-response functions: elaboration, estimation, and evaluation. Frontiers in systems neuroscience, 10, 109.

21. Naud, R., & Sprekeler, H. (2018). Sparse bursts optimize information transmission in a multiplexed neural code. Proceedings of the National Academy of Sciences, 115(27), E6329-E6338.

22. Ni, A. M., Ruff, D. A., Alberts, J. J., Symmonds, J., & Cohen, M. R. (2018). Learning and attention reveal a general relationship between population activity and behavior. Science, 359(6374), 463-465.

23. Nigam, S., Pojoga, S., & Dragoi, V. (2019). Synergistic Coding of Visual Information in Columnar Networks. Neuron, 104(2), 402-411.

24. Nienborg, H., & Cumming, B. G. (2006). Macaque V2 neurons, but not V1 neurons, show choice-related activity. Journal of Neuroscience, 26(37), 9567-9578.

25. Nienborg, H., & Cumming, B. G. (2009). Decision-related activity in sensory neurons reflects more than a neurons causal effect. Nature, 459(7243), 89.

26. Panzeri, S., Harvey, C. D., Piasini, E., Latham, P. E., & Fellin, T. (2017). Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. Neuron, 93(3), 491-507.

27. Pesaran, B., Pezaris, J. S., Sahani, M., Mitra, P. P., & Andersen, R. A. (2002). Temporal structure in neuronal activity during working memory in macaque parietal cortex. Nature neuroscience, 5(8), 805.

28. Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature, 497(7451), 585.

29. Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. Journal of neuroscience, 22(21), 9475-9489.

30. Saxena, S., & Cunningham, J. P. (2019). Towards the neural population doctrine. Current opinion in neurobiology, 55, 103-111.

31. Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. Annual review of neuroscience, 24(1), 1193-1216.

32. Tkačik, G., Marre, O., Amodei, D., Schneidman, E., Bialek, W., & Berry II, M. J. (2014). Searching for collective behavior in a large network of sensory neurons. PLoS computational biology, 10(1), e1003408.

33. Tkačik, G., Mora, T., Marre, O., Amodei, D., Palmer, S. E., Berry, M. J., & Bialek, W. (2015). Thermodynamics and signatures of criticality in a network of neurons. Proceedings of the National Academy of Sciences, 112(37), 11508-11513.

34. Shadlen, M. N., Britten, K. H., Newsome, W. T., & Movshon, J. A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. Journal of Neuroscience, 16(4), 1486-1510.

35. Shadlen, M. N., & Newsome, W. T. (1996). Motion perception: seeing and deciding. Proceedings of the national academy of sciences, 93(2), 628-633.

36. Shahidi, N., Andrei, A. R., Hu, M., & Dragoi, V. (2019). High-order coordination of cortical spiking activity modulates perceptual accuracy. Nature neuroscience, 22(7), 1148.

37. Silvanto, J., Cowey, A., Lavie, N., & Walsh, V. (2005). Striate cortex (V1) activity gates awareness of motion. Nature neuroscience, 8(2), 143.

38. Tajima, S., Yanagawa, T., Fujii, N., & Toyoizumi, T. (2015). Untangling brain-wide dynamics in consciousness by cross-embedding. PLoS computational biology, 11(11), e1004537.

39. Tajima, S., Mita, T., Bakkum, D. J., Takahashi, H., & Toyoizumi, T. (2017). Locally embedded presages of global network bursts. Proceedings of the National Academy of Sciences, 114(36), 9517-9522.

40. Uka, T., & DeAngelis, G. C. (2004). Contribution of area MT to stereoscopic depth perception: choice-related response modulations reflect task strategy. Neuron, 42(2), 297-310.

41. Wang, X. J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. Neuron, 36(5), 955-968.

42. Wang, X. J. (2008). Decision making in recurrent neuronal circuits. Neuron, 60(2), 215-234.

Supplementary material

S1 Fig.: Population signal

Population signal with learning and reconstruction on the *choice*; population signal with different length of the time window; population signal during the target stimulus.

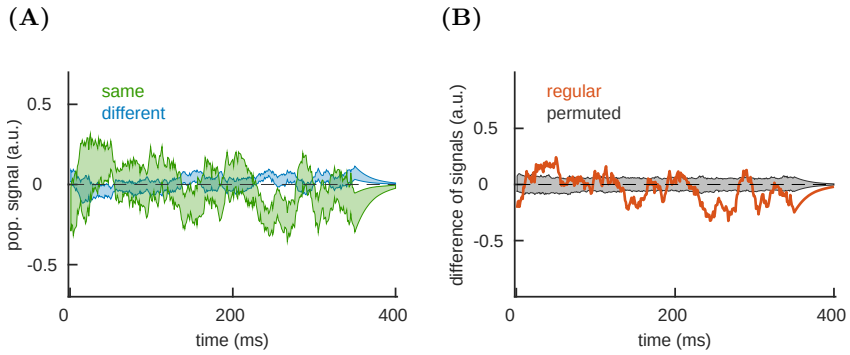

Figure S1.1. Learning and reconstruction on *choice* results in poor discrimination. (A) Population signal during test, where learning and reconstruction have used the information on *choice*. We show the mean ± SEM for the variability across sessions and show results for the choice "same" (green) and "different" (blue). **(B)** Session-averaged difference of population signals for the regular model (red) and the set of models with permuted class labels (gray). Parameters: $\lambda^{-1} = 20\ ms$, $N_{perm} = 1000$.

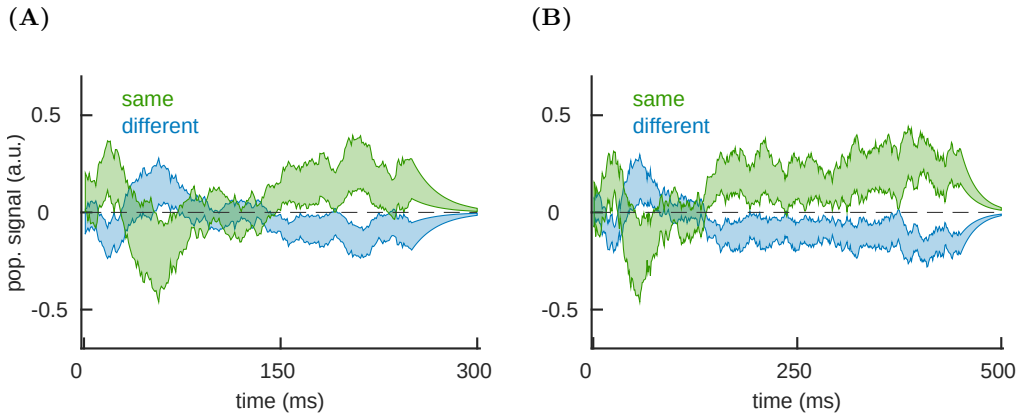

Figure S1.2. Population signal for different length of the time window. (A) Population signal during test for the length of the time window $K = 300$ ms. The plot shows the mean ± SEM across recording sessions for the behavioral choice "same" (green) and "different" (blue). The time window [0,K] ms w.r.t. stimulus onset is used for estimating the population vector as well as for computing the population signal. **(B)** Same as in **A**, but for the length of the time window $K = 500$. Parameter: $\lambda^{-1} = 20$ ms.

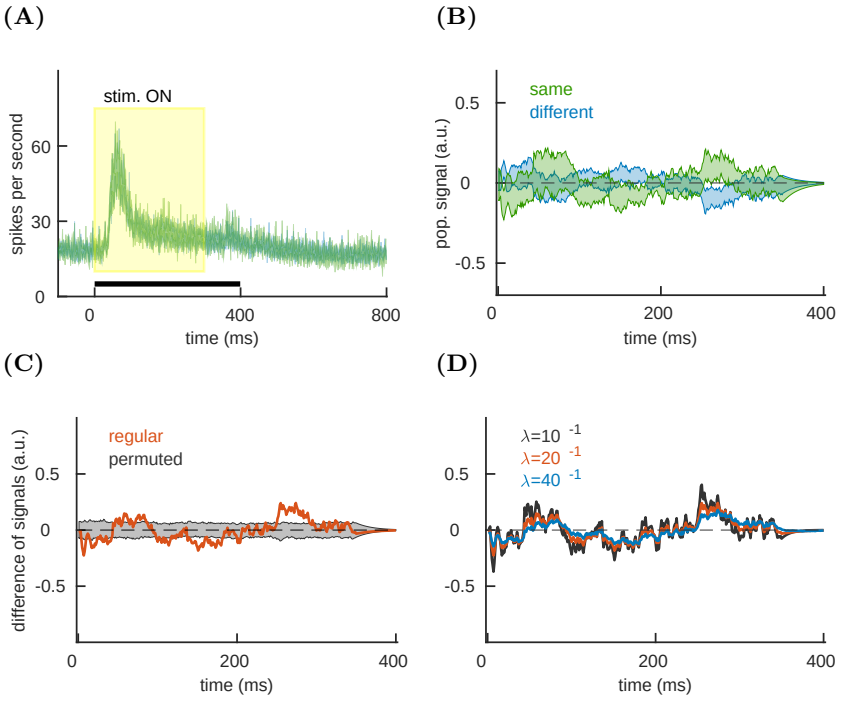

Figure S1.3. Behavioral choice cannot be predicted during visualization of the the target stimulus. (A) Population PSTHs during target for the choice "same" (green) and the choice "different" (blue). We show the mean ± SEM for the variability across recording sessions. The yellow window indicates the presence of the stimulus and the black bar indicates the time window, used for analysis ([0-400] ms with respect to the stimulus onset). **(B)** Population signal during target for the choice "same" (green) and "different" (blue). We show the mean ± SEM for the variability across sessions. Parameter: $\lambda^{-1} = 20\ ms$. **(C)** Session-averaged difference of population signals for the regular model (red) and the model with permuted class labels (gray). Parameters: $\lambda^{-1} = 20\ ms$, $N_{perm} = 1000$. **(D)** Session-averaged difference of the population signals for three values of the time constant of the convolution.

S2 Table: The synchrony between population signals for specified neuronal subpopulations

We report the correlation function at zero lag between population signals for specific neuronal subpopulations. The correlation function at zero lag measures the strength of synchrony between the two signals. The correlation function is implemented for population signals from the same trial (see methods). We measure the correlation function in single trials, then average across trials and across sessions.

$\lambda\ [ms^{-1}]$	$\lambda = 10$	$\lambda = 20$	$\lambda = 40$
plus & *minus*	-0.026	-0.029	-0.031

Table S2.1. Synchrony between population signals of *plus* and *minus* neurons

$\lambda\ [ms^{-1}]$	$\lambda = 10$	$\lambda = 20$	$\lambda = 40$
informative & uninformative	0.029	0.036	0.043

Table S2.2. Synchrony between population signals of informative and uninformative neurons

$\lambda\ [ms^{-1}]$	$\lambda = 10$	$\lambda = 20$	$\lambda = 40$
bursty & non-bursty	0.028	0.035	0.041

Table S2.3. Synchrony between population signals of bursty and non-bursty neurons

$\lambda\ [ms^{-1}]$ layers	$\lambda = 10$	$\lambda = 20$	$\lambda = 40$
SG & G	0.011	0.014	0.016
SG & IG	0.008	0.012	0.014
G & IG	0.007	0.010	0.013

Table S2.4. Synchrony between population signals for pairs of cortical layers

S3 Fig. Distance of decoding weights in the context of *stimulus+choice* and *choice*

The distance is computed for each single neuron and results are gathered across recording sessions. We then divide neurons according to a specific criterion (sign of the weight, burstiness, layers) and compare the groups with the two-tailed t-test.

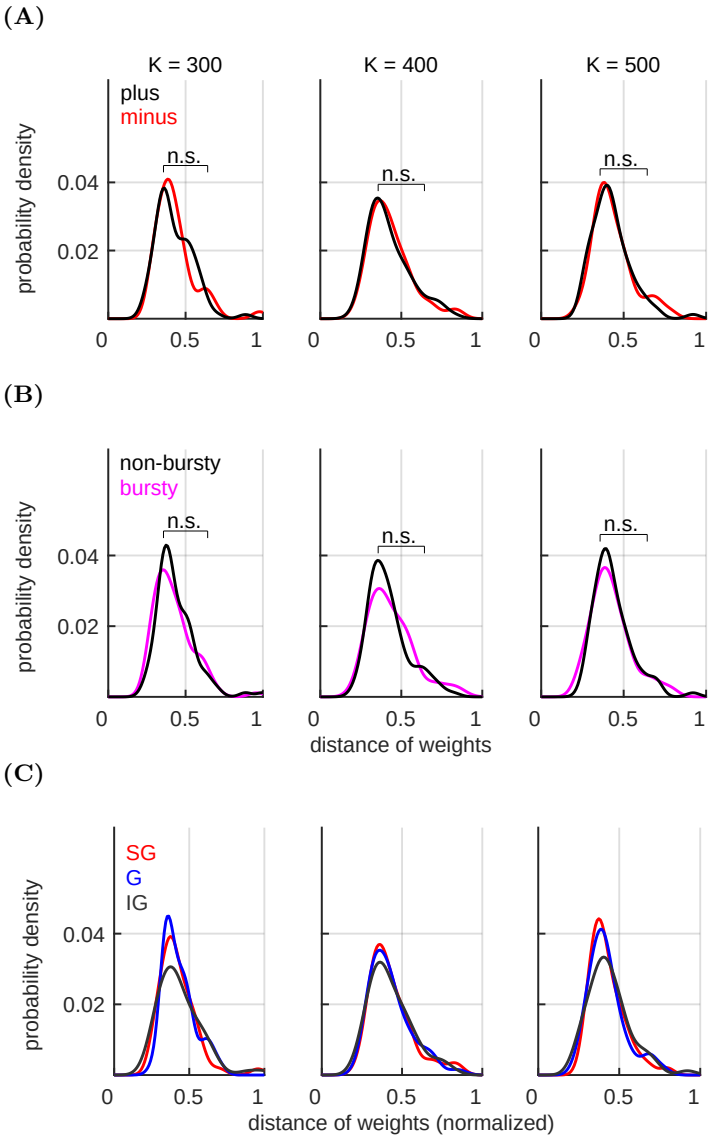

Figure S3.1. Sensitivity to informational context does not significantly differ between *plus* and *minus* neurons, between bursty and non-bursty neurons, nor across layers. (A) Distance of weights between w_n^{s+c} and w_n^c, collected across sessions and plotted as a distribution. We distinguish neurons with positive and negative weight and show results for the length of the time window $K = 300$ ms (left), $K = 400$ ms (middle) and $K = 500$ ms (right). **(B)** Same as in **A**, distinguishing bursty (magenta) and non-bursty (black) neurons. **(C)** Same as in **A**, distinguishing neurons in the SG (red), G (blue) and IG (black) layers.