PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# Inferring decoding strategies for multiple correlated neural populations

**Kaushik J. Lakshminarasimhan[1], Alexandre Pouget[2,3], Gregory C. DeAngelis[3], Dora E. Angelaki[1,4,5,6☯], Xaq Pitkow[1,4,7☯*]**

**1** Department of Neuroscience, Baylor College of Medicine, Houston, TX, United States of America, **2** Department of Basic Neuroscience, University of Geneva, Geneva, Switzerland, **3** Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY, United States of America, **4** Department of Electrical and Computer Engineering, Rice University, Houston, TX, United States of America, **5** Department of Mechanical and Aerospace Engineering, New York University, New York, United States of America, **6** Center for Neural Science, New York University, New York, United States of America, **7** Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, TX, United States of America

☯ These authors contributed equally to this work.
* xaq@rice.edu





**Citation:** Lakshminarasimhan KJ, Pouget A, DeAngelis GC, Angelaki DE, Pitkow X (2018) Inferring decoding strategies for multiple correlated neural populations. PLoS Comput Biol 14(9): e1006371. https://doi.org/10.1371/journal.pcbi.1006371

**Editor:** Frédéric E. Theunissen, University of California at Berkeley, UNITED STATES

**Received:** April 3, 2017

**Accepted:** July 17, 2018

**Published:** September 24, 2018



**Data Availability Statement:** All data files are available from the CRCNS database (http://dx.doi.org/10.6080/K07P8WKF).

**Funding:** This work was supported by NIH R01 DC04260, R21 DC014518, NSF NeuroNex 1707400, the Simons Collaboration for the Global Brain, and the Swiss National Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Abstract

Studies of neuron-behaviour correlation and causal manipulation have long been used separately to understand the neural basis of perception. Yet these approaches sometimes lead to drastically conflicting conclusions about the functional role of brain areas. Theories that focus only on choice-related neuronal activity cannot reconcile those findings without additional experiments involving large-scale recordings to measure interneuronal correlations. By expanding current theories of neural coding and incorporating results from inactivation experiments, we demonstrate here that it is possible to infer decoding weights of different brain areas at a coarse scale without precise knowledge of the correlation structure. We apply this technique to neural data collected from two different cortical areas in macaque monkeys trained to perform a heading discrimination task. We identify two opposing decoding schemes, each consistent with data depending on the nature of correlated noise. Our theory makes specific testable predictions to distinguish these scenarios experimentally without requiring measurement of the underlying noise correlations.

## Author summary

The neocortex is structurally organized into distinct brain areas. The role of specific brain areas in sensory perception is typically studied using two kinds of laboratory experiments: those that measure correlations between neural activity and reported percepts, and those that inactivate a brain region and measure the resulting changes in percepts. The two types of experiments have generally been interpreted in isolation, in part because no theory has been able combine their outcomes. Here, we describe a mathematical framework that synthesizes both kinds of results, giving us a new way to assess how different brain areas contribute to perception. When we apply our framework to experiments on behaving monkeys, we discover two models that can explain the perplexing finding that one

**Competing interests:** The authors have declared that no competing interests exist.

brain area can predict an animal's reported percepts, even though the percepts are not affected when that brain area is inactivated. The two models ascribe dramatically different efficiencies to brain computation. We show that these two models could be distinguished by a proposed experiment that measures correlations while inactivating different brain areas.

## Introduction

Although much is known about how single neurons encode information about stimuli, how neurons contribute to reported percepts is less well understood[1]. The latter, called the "decoding problem", seeks to identify how the brain uses the information contained in neuronal activity. Although some studies have sought to understand *principled* ways to decode population responses in the presence of correlated noise [2–12], the rules by which the brain *actually* integrates information across noisy neurons remain unclear.

Neuroscientists have traditionally investigated this question using two distinct approaches: causal or correlational. In causal approaches, experimenters selectively activate or inactivate brain regions of interest, and measure resulting perceptual or behavioural changes. In correlational approaches, experimenters measure correlations between behavioural choices and neuronal activity, typically quantified by 'choice probability' (reviewed in Ref. [13]) or, more straightforwardly, by 'choice correlation' (CC)[14,15]. If CCs reflect a functional link between neurons and behaviour, one would expect brain areas with greater CCs to contribute more strongly to behaviour. This naïve view is contradicted by recent results that reveal a striking dissociation between the magnitude of CCs and the effects of inactivation across brain systems in rodents[16,17] and primates[18,19]. In hindsight, this apparent disagreement is not all that surprising because the two techniques, on their own, yield results whose interpretation is fraught with major difficulties.

For instance, the CC of a neuron depends not only on its direct influence on behaviour but also on the influence of all the other neurons with which it is correlated. As an extreme example, a neuron that is not decoded at all could be correlated with one that is, and thus exhibit choice-related activity[9]. Recent theoretical results show that it is possible, in principle, to use knowledge of noise correlations to extract decoding weights from CCs[14]. However, directly measuring the correlational structures that matter for decoding may be extremely difficult[20]. This problem is compounded by the fact that behaviourally relevant information may be distributed across neurons in multiple brain areas, so neuronal CCs in one area may depend on activity in other areas. Moreover, in causal approaches, inactivation of one brain area could lead to a dynamic recalibration of decoding weights from other areas. Therefore, changes in behavioural thresholds following inactivation may not be commensurate with the contribution of the area.

When analysed in conjunction, however, results from correlational and causal studies may together provide constraints that can be used to precisely determine the relative contributions of the brain areas involved. In this work, we extend recent theories[14,15,20] and propose a general framework for inferring decoding weights of neurons across multiple brain areas using CCs and changes in behavioural threshold following inactivation. The two quantities together provide a direct estimate of the relative contributions of different areas without needing to precisely measure the correlation structure. This analysis is based on coarse-grained models of decoded neural noise that is correlated across populations. We demonstrate our technique by applying it to data from macaque monkeys trained to perform a heading discrimination task. In this task, there is a known discrepancy[18,21–23] between CCs and the effects of inactivating two brain areas: although neurons in the ventral intraparietal (VIP) area were found to be substantially

better predictors of the animal's choices than dorsal medial superior temporal (MSTd) neurons, performance is impaired by inactivating MSTd but not VIP. We use our framework to extract key properties of the decoder that can account for these counter-intuitive results. To our surprise, we find that, depending on the structure of correlated noise, experimental data are consistent with two opposing schemes that attribute either too much or too little weight to VIP. We use our theory to make specific testable predictions to distinguish these schemes using CCs measured during inactivation, again without measuring the detailed noise correlations.

## Results

Our framework for understanding neural decoding involves three main ingredients: an analysis of choice correlations and discrimination thresholds, two classes of models for noise correlations with different information content, and coarse-grained descriptions of those models for multiple populations. Our analysis proceeds as follows. We begin in section **Decoding framework** with some core definitions for neural population responses and estimation tasks based on decoding from multiple populations. Then, in the section **Analysis of choice correlations**, we describe the expected patterns of choice-related activity under the assumptions of optimal and suboptimal decoding. These patterns depend on the structure of neural noise, so in the section, **Models of neural variability**, we next describe two fundamentally different noise models, whose information content is extensive (*i.e*. growing with population size) or limited. We then refine these models for multiple populations in the section **Coarse-grained noise models for multiple populations**. Next we return to choice correlations to explore consequences of this coarse-grained description in the section **Coarse-grained choice correlations**. Our general theoretical analysis concludes in **Combining choice correlations and inactivation effects to infer decoding of distinct populations**. Finally, we specialize this theory to two populations as we apply it to experimental data.

Some readers wishing to skip some of the mathematical details may wish to read the sections **Decoding framework**, which sets out the basic concepts we invoke, and **Models of neural variability**, which describes the two main noise models we contrast, before jumping to **Application to neural data**.

### Decoding framework

We consider a linear feedforward network in which the firing rates $\mathbf{r} = [r_1,\ldots,r_N]$ of the $N$ neurons are tuned to the stimulus $s$ as $\mathbf{f}(s) = \langle \mathbf{r}|s\rangle$, where the angle brackets denote an average over trials conditioned on the stimulus. The responses on a single trial differ from their averages by some noise with variance $\sigma_k^2$ for neuron $k$, and exhibit a covariance $\Sigma = \langle \mathbf{r}\mathbf{r}^{\mathrm{T}}|s\rangle - \mathbf{f}(s)\mathbf{f}(s)^{\mathrm{T}}$ that we assume is stimulus-independent. These neural responses are combined linearly using weights $\mathbf{w}$ to yield a locally unbiased estimate $\hat{s}$ of the stimulus according to $\hat{s} = \mathbf{w}^{\mathrm{T}}(\mathbf{r} - \mathbf{f}(s_0)) + s_0$. Here *local* means that the stimulus is near a reference $s_0$, which we will now take to be 0 without loss of generality, and $\mathbf{f}(s_0)$ is the mean population response to that reference. *Unbiased* estimation means that the estimate is accurate on average, so that $\langle \hat{s}|s\rangle = s$. In the experiments we model, the animals indeed are unbiased after training.

The performance of a decoder is often characterized by the variance $\varepsilon$ of its estimate:

$$\varepsilon = \langle \hat{s}^{\,2}\rangle - \langle \hat{s}\rangle^2 = \langle (\mathbf{w}^{\mathrm{T}}\mathbf{r})^2\rangle - (\mathbf{w}^{\mathrm{T}}\mathbf{f})^2 = \mathbf{w}^{\mathrm{T}}\Sigma\mathbf{w} \tag{1}$$

Other common measures of performance are the discrimination threshold $\vartheta$, sensitivity index, $d'$, and Fisher information $J$. These measures are all closely related. We will often refer to the discrimination threshold $\vartheta$, which is the stimulus difference, $\Delta s$, required for reliable binary discrimination between two categories when discrimination is based on an estimator with finite variance.

When 'reliable' is 68% correct, then this threshold is just the estimate's standard deviation, $\vartheta = \sqrt{\varepsilon}$. This definition coincides with the sensitivity index $d' = \Delta\mu/\sigma_{\hat{s}} = 1$, when the mean difference, $\Delta\mu$, between estimates for the two stimuli is the same size as the standard deviation, $\sigma_{\hat{s}}$, of those estimates. When the neural response mean $\mathbf{f}(s)$ is tuned to the stimulus, but other statistics do not provide additional information (*i.e.* for responses drawn from the exponential family), then the Fisher information, $J$, is exactly equal to the inverse variance of an unbiased, locally optimal linear estimator: $J = 1/\varepsilon$ (also assuming differentiable tuning curves and non-singular noise covariance).

Many experiments assess performance using a two-alternative forced-choice experiment (2AFC). They quantify performance by the discrimination threshold, $\vartheta$, which is the stimulus difference required for reliable binary discrimination (68% correct) (see **Methods**), and assess neural decoding based on choice probabilities[24]. However, theoretical results about decoding are much simpler when applied to continuous estimation (which we will consider to be a continuous 'choice'). Conveniently, local continuous estimation and fine discrimination are closely related. For example, as mentioned above, the discrimination threshold $\vartheta$ is equal to the standard deviation of an unbiased local estimator, $\sigma_{\hat{s}}$, if the output variability is Gaussian. Under the same assumptions, choice correlation has a simple near-affine relation to choice probability (see **Methods**, [15]). We thus first describe the theory in terms of a local estimation task, and later apply the suitable transformations when we analyze data from binary discrimination tasks.

If the brain decodes signals linearly from multiple populations of neurons, its overall estimate $\hat{s}$ can always be expressed as a linear combination of unbiased estimates from each population separately:

$$\hat{s} = \mathbf{a}^{\mathrm{T}}\hat{\mathbf{s}} \tag{2}$$

where $\hat{\mathbf{s}} = [\hat{s}_1, \ldots, \hat{s}_Z]$ is a vector of separate estimates from each of $Z$ populations, and $\boldsymbol{a}$ is a vector of *scaling factors* for each estimate to create one overall estimate. We call these 'scaling factors' to distinguish them from the weights given to individual neurons. Thus the problem of decoding multiple populations can be viewed as one of scaling and combining estimates from individual populations. Note that this is equivalent to a single linear decoder of all populations together using $\mathbf{w} = [a_1\mathbf{w}_1 \cdots a_Z\mathbf{w}_Z]$.

For locally linear decoding, the assumption of no bias implies a normalization constraint on the weights and scaling factors. An unbiased estimate should match the stimulus, on average; and so a change in the estimate should match the change in the stimulus, on average: $\partial_s\langle\hat{s}|s\rangle = \partial_s\mathbf{w}^T(\mathbf{f}(s) - \mathbf{f}(0)) \approx \mathbf{w}^{\mathrm{T}}\mathbf{f}'(s) = \partial_s s = 1$. Analogously, unbiased scaling factors of individually unbiased estimates $\hat{s}_z$ satisfy $\mathbf{a}^{\mathrm{T}}\partial_s\langle\hat{\mathbf{s}}|s\rangle = \mathbf{a}^T\mathbf{1} = 1$, where $\mathbf{1}$ is a vector of all ones and where each population estimate $\hat{s}_x = \mathbf{w}_x^T(\mathbf{f}_x(s) - \mathbf{f}_x(0))$ obeys the normalization $\mathbf{w}_x^T\mathbf{f}_x'(s) = 1$.

Using this decomposition into populations, we can dissociate how the weight *patterns* within each subpopulation ($\mathbf{w}_x$) and their *scaling factors* ($a_x$) affect the output of the decoder. This mathematical separation is also appealing because it provides a common framework to synthesize results from experiments conducted at two fundamentally different levels of granularity. One class of experiments involves making fine measurements such as the correlation between trial-by-trial fluctuations in the activity $r_k$ of an individual neuron $k$ and the animal's decision (**Fig 1A**). The second class of experiments studies causation by measuring behavioural effects of inactivating certain candidate brain areas. For perceptual discrimination tasks, this is done by comparing coarse measures such as the animal's behavioural performance before ($\vartheta$) and after ($\vartheta_{-x}$) inactivating population $x$ (**Fig 1B**).

We would like to use these experimental measurements to identify the relative behavioural contributions of various brain areas. Therefore we will present a technique to infer neuronal

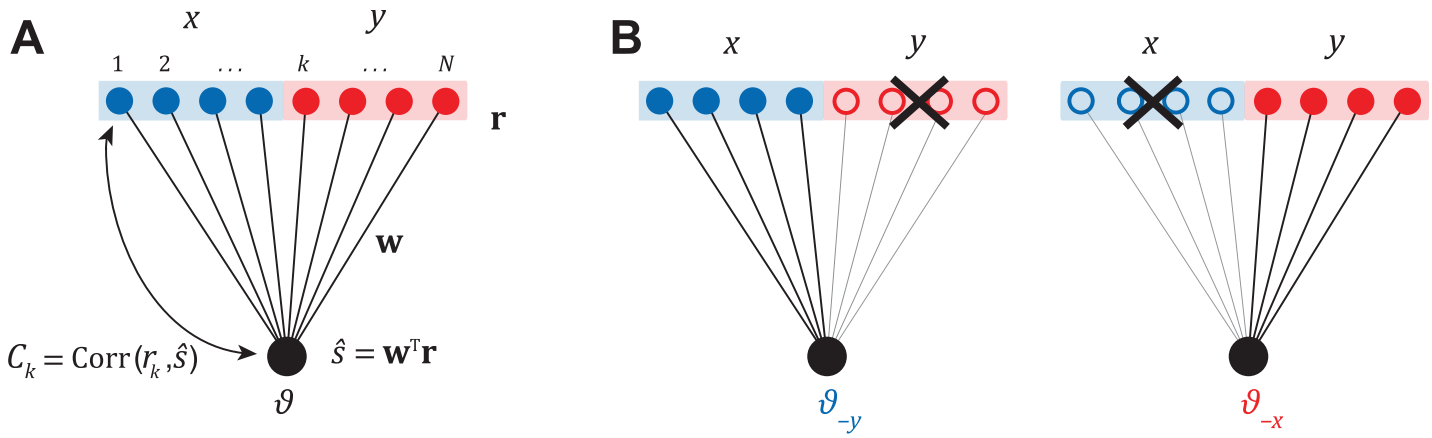


**Fig 1. Experimental strategies. (A)** An illustration of a feedforward network with linear readout. The decoder linearly combines the activity **r** of neurons in two populations *x* and *y* with weights **w**, to produce an estimate $\hat{s}$ of the stimulus. Activity of individual neurons $r_k$ is correlated with $\hat{s}$ and is quantified by either the choice probability $CP_k$, or the closely related choice correlation $C_k$. In an optimal system, the weights **w** generate choice correlations that satisfy **Eq (4)**. **(B)** In inactivation experiments, the neurons from each population are inactivated and the resulting changes in behavioural threshold are recorded.

https://doi.org/10.1371/journal.pcbi.1006371.g001

readout weights in multiple brain areas, focusing primarily on how to extract the scaling factors, $a_x$, of the brain areas rather than the fine structures, $\mathbf{w}_x$, of their decoding weights.

## Analysis of choice correlations

Choice correlation of a neuron *k* is defined as the correlation coefficient between its response $r_k$ and the animal's estimate of the stimulus $\hat{s}$, $C_k = \mathrm{Corr}(r_k, \hat{s}|s)$, across repeated trials with the same stimulus *s*. Substituting the estimate into this correlation, we find:

$$C_k = \frac{\mathrm{Cov}(r_k, \mathbf{r}^T\mathbf{w}|s)}{\sqrt{\mathrm{Var}(r_k|s)\mathrm{Var}(\mathbf{r}^T\mathbf{w}|s)}} = \frac{\langle r_k\mathbf{r}^T\mathbf{w}\rangle - \langle r_k\rangle\langle\mathbf{r}^T\mathbf{w}\rangle}{\sqrt{\sigma_k^2(\langle\mathbf{w}^T\mathbf{r}\mathbf{r}^T\mathbf{w}\rangle - \langle\mathbf{w}^T\mathbf{r}\rangle\langle\mathbf{r}^T\mathbf{w}\rangle)}} = \frac{(\Sigma\mathbf{w})_k}{\sqrt{\sigma_k^2\mathbf{w}^T\Sigma\mathbf{w}}} \quad (3)$$

where the noise variance for neuron *k* is $\mathrm{Var}(r_k|s) = \sigma_k^2 = \Sigma_{kk}$. All neurons' choice correlations can then be expressed together in vector form as $\mathbf{C} = \frac{S^{-1}\Sigma\mathbf{w}}{\sqrt{\mathbf{w}^T\Sigma\mathbf{w}}}$, where *S* is a diagonal matrix of the standard deviations.

These choice correlations follow a particularly simple pattern if readout weights are locally optimal [15] as obtained from linear regression as $\mathbf{w} \propto \Sigma^{-1}\mathbf{f}'$. If we substitute these optimal weights into **Eq (3)**, the inverse covariance from the weights cancels the covariance driving the choice correlations:

$$\begin{aligned} C_{k,\mathrm{opt}} &= \frac{(\Sigma\Sigma^{-1}\mathbf{f}')_k}{\sqrt{(\Sigma^{-1}\mathbf{f}')^T\Sigma(\Sigma^{-1}\mathbf{f}')\sigma_k^2}} \\ &= \frac{f_k'}{\sigma_k}\frac{1}{\sqrt{\mathbf{f}'^T\Sigma^{-1}\mathbf{f}'}} \\ &= \frac{\vartheta}{\vartheta_k} \end{aligned} \quad (4)$$

where $C_{k,\mathrm{opt}}$ is the choice correlation of neuron *k* expected from optimal decoding, $\vartheta_k = f_k'/\sigma_k$ is the discrimination threshold of neuron *k* (or, equivalently, the standard deviation of an unbiased estimator based only on that neuron's response), and $\vartheta$ is the behavioural discrimination threshold. If decoding were optimal, then this behavioural threshold will match the

standard deviation of a locally optimal unbiased estimator based on the whole population, $\vartheta = (\mathbf{f}'^T\Sigma^{-1}\mathbf{f}')^{-1/2}$. By itself, such a match would be strong evidence for optimal decoding, but testing this would require recording from all relevant neurons in the brain. The relationship in **Eq (4)** is thus a far more practical test for optimal decoding.

If all neurons from multiple populations satisfy the above equation, this gives us strong evidence that the neuronal weights — and consequently also the relative scaling factors $\boldsymbol{a}$ of different populations — are optimal. As we will see later, the exact values of $\boldsymbol{a}$ can then be directly extracted from the behavioural thresholds following inactivation of those areas.

The pattern of choice correlations generated by any generic *sub*optimal decoder is more complicated, as it depends explicitly on the structure of noise covariance and the readout weights [14]. For a population of $N$ neurons, the noise covariance $\Sigma$ describes, for a fixed stimulus, the power along $N$ orthogonal modes of variation. Each of these modes could contribute to the overall choice correlation, depending on how strongly that mode is decoded. We express the decoding weights of a suboptimal decoder in terms of the covariance, as $\mathbf{w} = (\Sigma^{-1}\mathbf{g})/\mathbf{f}'^{\mathrm{T}}\Sigma^{-1}\mathbf{g}$ where $\mathbf{g}$ could be any vector in $\mathbb{R}^N$. The normalization ensures that this decoder is locally unbiased, satisfying $\mathbf{w}^T\mathbf{f}' = 1$.

$$\mathbf{C} = \frac{S^{-1}\Sigma\mathbf{w}}{\sqrt{\mathbf{w}^{\mathrm{T}}\Sigma\mathbf{w}}} = \frac{S^{-1}\Sigma\mathbf{w}}{\vartheta} = \frac{S^{-1}}{\vartheta\mathbf{f}'^{\mathrm{T}}\Sigma^{-1}\mathbf{g}}\mathbf{g} \tag{5}$$

Note that this recovers the optimal expression given by equation (4) if $\mathbf{g}$ is replaced by $\mathbf{f}'$. We now rewrite $\mathbf{g}$ in the basis of the eigenmodes $\mathbf{u}^i$ of the covariance $\Sigma$, using $\mathbf{g} = \sum_{i=1}^{N}\mathbf{u}^i\mathbf{u}^{i^T}\mathbf{g}$. By multiplying and dividing by $\mathbf{u}^{i^T}\mathbf{f}\prime$, we can decompose the choice correlations for a suboptimal decoder into a weighted combination of optimal choice correlations patterns $\mathbf{C}^i_{\mathrm{opt}}$ arising from each eigenmode:

$$\begin{aligned}\mathbf{C} &= \frac{S^{-1}}{\vartheta\mathbf{f}'^{\mathrm{T}}\Sigma^{-1}\mathbf{g}}\sum_{i=1}^{N}\mathbf{u}^i\mathbf{u}^{i^{\mathrm{T}}}\mathbf{g} \\ &= \frac{S^{-1}}{\vartheta\mathbf{f}'^{\mathrm{T}}\Sigma^{-1}\mathbf{g}}\sum_{i=1}^{N}\mathbf{u}^i(\mathbf{u}^{i^T}\mathbf{f}')\frac{(\mathbf{u}^{i^{\mathrm{T}}}\mathbf{g})}{(\mathbf{u}^{i^{\mathrm{T}}}\mathbf{f}')} \\ &= \sum_{i=1}^{N}\beta_i\mathbf{C}^i_{\mathrm{opt}}\end{aligned} \tag{6}$$

where

$$\mathbf{C}^i_{\mathrm{opt}} = \vartheta\, S^{-1}\mathbf{u}^i(\mathbf{u}^{i^T}\mathbf{f}') \tag{7}$$

$\mathbf{C}^i_{\mathrm{opt}}$ is essentially the $i$'th noise mode $\mathbf{u}^i$ rescaled by the individual neural sensitivity, and $\beta_i = \frac{1}{\vartheta^2(\mathbf{f}'^{\mathrm{T}}\Sigma^{-1}\mathbf{g})}\frac{(\mathbf{g}^{\mathrm{T}}\mathbf{u}^i)}{(\mathbf{f}'^{\mathrm{T}}\mathbf{u}^i)}$. These multipliers $\beta_i$ reflect the extent of suboptimality. When decoding weights are optimal, then the readout direction (again in units of the covariance) is $\mathbf{g} = \mathbf{f}'$, leading to $\beta_i = 1$ for all $i$. Thus, for optimal decoding the above equation reduces to **Eq (4)**.

In principle, elements of $\beta_i$, and thus properties of the decoding weights, can be estimated by regressing measured choice correlations against individual columns of the matrix of choice correlations $C_{\mathrm{opt}}$ predicted by optimal decoding. In practice, it is very difficult to estimate all of the multipliers $\beta_i$ because the components $C^i_{k,\mathrm{opt}}$ depend on the individual noise modes of $\Sigma$ (**Eq (7)**). Directly measuring $\Sigma$ is a notoriously challenging task [20] that involves simultaneously recording the activity of a large population of neurons, and is nearly impossible for certain areas due to the geometry of the brain. Even if such recordings could be performed, it

would be challenging to get an accurate assessment of the fine structure of the covariance with limited data, since the number of parameters to measure increases with population size faster than the number of measurements. Fortunately, since neuronal choice correlations are measurably large, it follows that one can infer the animal's decoding weights with reasonable precision by estimating the few leading multipliers that depend only on the most dominant modes of covariance. This is because if the correlated noise modes with small variance were to dominate the decoder, then only a tiny fraction of each neuron's variations would propagate to the decision, leading to immeasurably small choice correlations[15] (**S1 Fig**). It is possible to model properties of the leading modes of covariance without large-scale recordings, and we will consider two different noise models: *extensive information* and *limited information*.

## Models of neural variability

**Extensive information model.** A common way to measure important components of the covariance structure is through pairwise recordings. Noise covariance measured between pairs of neurons can be modeled as a function of their response properties, such as the difference in their preferred stimulus or the similarity of their tuning functions, to obtain empirical models of noise.

One such model is limited-range noise correlations[25–30], so called because they are proportional to signal correlation and thereby limited in range to pairs with similar tuning. We use this model to approximate a full noise covariance for all neurons in the population[31,32]. Specifically, we assume that the typical noise correlation coefficient $\bar{R}_{ij}$ between responses of two neurons $i$ and $j$ is given by

$$\bar{R}_{ij} = (1 - m)\delta_{ij} + mR_{ij}^{\text{sig}} \tag{8}$$

where $R_{ij}^{\text{sig}} = \text{Corr}(f_i, f_j)$ is the signal correlation, i.e. the correlation coefficient between neurons' mean responses over a uniform distribution of stimuli $s$ and the proportionality $m$ between signal and noise correlations can be empirically determined (see **Methods**). To match Poisson-like properties of neural responses, model variances are set equal to the mean responses, and this scaling produces a covariance of $\Sigma_{ij} = R_{ij}\sqrt{f_i f_j}$. This has been a common noise model in the study of population codes[25–30]. Although the resulting covariance matrix is unlikely to capture fine details accurately, if the model is reasonable then most of the variance would be captured by the leading modes.

In an extensive information model, the amount of information encoded by the neural activity grows with population size [33–35], hence the name. If the brain extracts information by a decoder restricted only to the noisiest subspace given by these leading noise modes, this would recover just a tiny fraction of the total available information. Although this is radically suboptimal, this is the only way an extensive information model can explain the large magnitude of neuronal choice correlations[15].

**Limited information model.** Extensive information models are based on measurements of neural populations but, as we mentioned above, current recordings are not sufficient to measure or even infer the covariance matrix *in vivo*. It is therefore possible that information in cortex is not extensive. Indeed, the extensive information model conflicts with the fact that cortical neurons receive their inputs from a smaller population of neurons. The cortex must then inherit not only the input signal but also any noise in that input. This generates information-limiting correlations [15,20] in the cortex, a form of correlated noise that looks exactly like the signal and thus cannot be averaged away by adding more cortical neurons. Since inferring the brain's decoding weights from choice-related activity depends on the noise covariance, we also consider the consequences of information-limiting correlations.

For fine discrimination between two neighboring stimuli $s$ and $s + \delta s$, the signal is given by the change in mean population responses $\mathbf{f}(s + \delta s) - \mathbf{f}(s) \approx \delta s\, \mathbf{f}'(s)$. Information-limiting correlations for this task thus fluctuate along the direction $\mathbf{f}'$, generating a covariance containing differential correlations [20] — that is, a covariance component proportional to $\mathbf{f}'\mathbf{f}'^{\mathrm{T}}$. The constant of proportionality, which we denote as $\varepsilon$, represents the variance of information-limiting correlations. According to this model, the total noise covariance $\Sigma_{\mathrm{IL}}$ for the information-limiting model can be decomposed into a general noise covariance $\Sigma$ (which we assume follows the extensive information model) and the information-limiting component:

$$\Sigma_{\mathrm{IL}} = \Sigma + \varepsilon \mathbf{f}'\mathbf{f}'^{\mathrm{T}} \tag{9}$$

The variance of a locally optimal linear estimator based on a neural population with this noise covariance is given by [20]:

$$\langle \delta\hat{s}^{\,2} \rangle = [\mathbf{f}'^{\mathrm{T}} \Sigma_{\mathrm{IL}}^{-1} \mathbf{f}']^{-1} = (\mathbf{f}'^{\mathrm{T}} \Sigma^{-1} \mathbf{f}')^{-1} + \varepsilon \approx \varepsilon \tag{10}$$

where we have used the Sherman-Morrison lemma to invert $\Sigma_{\mathrm{IL}}$. The estimator variance due to the extensive information term $(\mathbf{f}'^{\mathrm{T}} \Sigma^{-1} \mathbf{f}')^{-1}$ shrinks with population size [20,33,34], and is eventually dominated by the information-limiting noise variance $\varepsilon$. With increasing population size, both the signal $\mathbf{f}'$ and the information-limiting component $\varepsilon \mathbf{f}'\mathbf{f}'^{\mathrm{T}}$ grow identically, eventually resulting in no further improvement in signal-to-noise ratio, and thus no improvement in discriminability. In general, $\varepsilon$ could be very small, and hence information-limiting correlations may be very hard to detect with limited data as they are easily swamped by noise arising from other sources. Nevertheless, this noise has enormous implications for decoding large populations because it limits the total information to $1/\varepsilon$.

## Coarse-grained noise models for multiple populations

In this section we describe these two noise correlation models coarsely, at the population level, so that we can use the shared fluctuations between populations to reveal the decoder's scaling factors. To attribute scaling factors to each of $Z$ decoded populations, one must consider at least $Z$ modes of the noise covariance, one per population. We will restrict our attention to decoders inhabiting only these leading modes. If there are $Z$ dominant noise modes and they are correlated across populations, then we can approximate $\Sigma$ with a rank-$Z$ noise covariance matrix composed of both independent and correlated noise between the populations.

**Multi-population limited information model.** When dealing with multiple populations (e.g., in different brain areas), one has to keep in mind that although they may together receive limited information, they need not inherit it from exactly the same upstream neurons. Therefore, we construct a more general model allowing the different populations to receive both distinct and shared information. To describe this, we separate a low-rank information-limiting fluctuations from a general noise covariance $\Sigma$ (which we assume follows the extensive information model),

$$\Sigma_{\mathrm{IL}} = \Sigma + FEF^{\mathrm{T}} \tag{11}$$

Here $F$ is an $N \times Z$ block-diagonal matrix

$$F = \begin{pmatrix} \boldsymbol{f}'_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \boldsymbol{f}'_Z \end{pmatrix} \tag{12}$$

and $\mathbf{f}'_z$ is a vector of stimulus sensitivities for all neurons in population $z$, with elements $f'_{iz}$, and $E$ is a $Z \times Z$ covariance for information-limiting noise in each population. The covariance between two neurons in this more general information-limiting model would still be proportional to the product of the derivative of their tuning curves. However, the constant of proportionality varies depending on whether the neurons are both from the same population $x$ ($E_{xx}$), both from $y$ ($E_{yy}$), or from different populations ($\varepsilon_{xy}$):

$$\Sigma_{\mathrm{IL}} = \Sigma + \begin{pmatrix} \varepsilon_{xx}\boldsymbol{f}'_x\boldsymbol{f}'^{\mathrm{T}}_x & \varepsilon_{xy}\boldsymbol{f}'_x\boldsymbol{f}'^{\mathrm{T}}_y & \cdots \\ \varepsilon_{xy}\boldsymbol{f}'_x\boldsymbol{f}'^{\mathrm{T}}_y & \varepsilon_{yy}\boldsymbol{f}'_y\boldsymbol{f}'^{\mathrm{T}}_y & \cdots \\ \cdots & \cdots & \ddots \end{pmatrix} \tag{13}$$

Analogous to the information-limiting noise variance $\varepsilon$ in the single population case (**Eq (10)**), elements of $E$ once again determine the variance of the locally optimal linear estimators (and thus optimal discrimination thresholds) for individual populations, as well as for all populations together (**S2 Text**). We call the noise $\varepsilon_{xx}\boldsymbol{f}'_x\boldsymbol{f}'^{T}_x$ in each population $x$ "*locally* information-limiting noise" because it is local to one population $x$. For large populations with this noise structure, the total information content within population $x$ alone is limited to $1/\varepsilon_{xx}$.

By itself, this local noise does not guarantee that the complete population is globally information-limited: that depends on how the noise in different populations is correlated. For example, input from another brain area might add some locally information-limiting noise [36], which could in principle be removed again by appropriately decoding both brain areas together. Depending on the covariance between information-limiting noise across populations, $\varepsilon_{xy}$, different populations may contain completely redundant, independent, or synergistic information [37,38]. However, the information in all populations together may be limited as well, ultimately by the $\boldsymbol{f}'\boldsymbol{f}'^{T}$ component of the covariance $\Sigma$. We call this component "globally information-limiting noise".

Correlations that limit information also cause redundancy. As a consequence, many different decoding weights extract essentially the same information. The population is then robust to some amount of suboptimal decoding, which makes it easier to achieve near-optimal behavioural performance [15]. In the locally information-limited noise model for multiple populations described above, this robustness also holds within each population individually. In this case, a separate decoder for each population $x$ produces an estimate $\hat{s}_x$ that is near-optimal for the corresponding areas. Importantly, however, these estimates may have different variances, and may even covary, so they need to be properly combined to produce a good single estimate according to **Eq (2)**. While information-limiting correlations within each area would make the system generally robust to the choice of weight patterns $\mathbf{w}_x$, suboptimality could yet arise from an incorrect scaling $a_x$ of each individually near-optimal estimate. This is because after the dimensionality reduction from large redundant populations down to a single unbiased estimate per population, most of the redundancy has been squeezed out: just one degree of freedom remains for the decoder, so different ways of combining the estimates are not equivalent.

**Multi-population extensive information model.** For the extensive information model, we can also define a useful rank-$Z$ approximation of the relevant components of the noise covariance $\Sigma$. Let $\mathbf{u}_x$ denote the leading eigenvector of population $x$'s covariance $\Sigma_{xx}$, with corresponding eigenvalue $\lambda_x$. Note that these are not the eigenvectors of the full covariance matrix, just of the covariances for each population separately. If, in the full covariance, the leading modes of different populations $\boldsymbol{x}$ and $\boldsymbol{y}$ interact to produce correlated noise with strength $\lambda_{xy}$, then we approximate the full covariance by $\Sigma = ULU^{\mathrm{T}}$ where, analogously with

Eq (12),

$$U = \begin{pmatrix} \mathbf{u}_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{u}_Z \end{pmatrix} \tag{14}$$

and the $Z \times Z$ matrix

$$L = \begin{pmatrix} \lambda_1 & \cdots & \lambda_{1Z} \\ \vdots & \ddots & \vdots \\ \lambda_{1Z} & \cdots & \lambda_Z \end{pmatrix} \tag{15}$$

In the extensive information model, an optimal decoder would largely avoid the largest noise modes. However, optimal decoding of the extensive model is thoroughly ruled out by experimental measurements described below (see section 'Test for Optimality'). Thus, for our coarse-grained multi-population model, we assume the brain's decoder is limited to the noisiest mode for each population, while it has complete freedom to combine estimates derived thusly from each population. Future refinements of this coarse-grained framework could consider decoding other modes per population instead, or more modes.

Unlike elements of information-limiting noise $E$ in **Eq (13)**, elements of $L$ cannot be directly related to the variance of the output estimator $\hat{s}$ because the latter depends not only on the magnitude of noise ($\lambda_x$) but also on the signal ($\mathbf{u}_x^{\mathrm{T}}\mathbf{f}'_x$). But we can rescale each element of $L$ to obtain $E$, and express a low-rank approximation of the covariance $\Sigma$ in terms of $E$ as:

$$\Sigma = U(U^{\mathrm{T}}F)E(U^{\mathrm{T}}F)^{\mathrm{T}}U^{\mathrm{T}} \tag{16}$$

where $E = (U^{\mathrm{T}}F)^{-1}L(U^{\mathrm{T}}F)^{-1}$, so the elements of $E$ are related to $L$ as: $\varepsilon_{xx} = \frac{\lambda_x}{(\mathbf{u}_x^{\mathrm{T}}\mathbf{f}'_x)^2}$ and $\varepsilon_{xy} = \frac{\lambda_{xy}}{(\mathbf{u}_x^{\mathrm{T}}\mathbf{f}'_x)(\mathbf{u}_y^{\mathrm{T}}\mathbf{f}'_y)}$. Just like the case of information-limiting noise, the elements of $E$ again determine optimal thresholds according to **S2 Text** (Eqn (S2.1) – (S2.2)), but with one key distinction: whereas those thresholds correspond to the output of optimal decoding for each population in the case of information-limiting noise, these correspond to outputs of optimal decoding only within the subspace of the $Z$ populations' leading modes in the case of extensive information model. Note that we can use the formulation in **Eq (16)** to derive information-limiting noise (**Eq (11)**) as a special case by using $\mathbf{u}_x = \mathbf{f}'_x/\|\mathbf{f}'_x\|$ to recover $\Sigma = FEF^{\mathrm{T}}$.

## Coarse-grained choice correlations

These coarse-grained representations of population variability reflect the dominant decoded mode in each population. This level of description allows us to focus on how information is combined between populations. If the brain indeed combines activity from different areas suboptimally, then simplifying **Eq (6)** in the presence of information-limiting correlations gives choice correlations within each area that are not equal to the optimal choice correlations, but

are still proportional to them.

$$\begin{aligned}\mathbf{C} &= \frac{S^{-1}\Sigma\mathbf{w}}{\sqrt{\mathbf{w}^{\mathrm{T}}\Sigma\mathbf{w}}} \approx \frac{S^{-1}FEF^{\mathrm{T}}\mathbf{w}}{\sqrt{\mathbf{w}^{\mathrm{T}}\Sigma\mathbf{w}}} \approx \frac{S^{-1}FE\mathbf{a}}{\sqrt{\mathbf{a}^{\mathrm{T}}E\mathbf{a}}} \\ &= \frac{E\mathbf{a}}{\mathbf{a}^{\mathrm{T}}E\mathbf{a}}\sqrt{\mathbf{a}^{\mathrm{T}}E\mathbf{a}}(S^{-1}F) = \frac{E\mathbf{a}}{\mathbf{a}^{\mathrm{T}}E\mathbf{a}}\frac{\vartheta}{\vartheta_k} \\ &= \beta\frac{\vartheta}{\vartheta_k}\end{aligned} \tag{17}$$

where $\beta_x = \frac{(E\mathbf{a})_x}{\mathbf{a}^{\mathrm{T}}E\mathbf{a}}$. Under conditions of suboptimality, choice correlations in different brain areas $x$ may have different multipliers $\beta_x$ which depend on the scaling of the brain areas and on the covariance between the estimates $\hat{s}_x$ that can be derived from them. These multipliers $\beta_x$ can be directly identified by regressing measured choice correlations against $\vartheta/\vartheta_k$, the choice correlations predicted for optimal decoding. S4 Text shows that a similar relation holds for the extensive information model when only the leading mode of each population is decoded (S4 Text – Eqn (S4.1)).

## Combining choice correlations and inactivation effects to infer decoding of distinct populations

In the previous section, we showed how to reduce the fine structure of choice correlations down to one number for each population, the slope $\beta_x$ of its choice correlation. We will now show how these multipliers can be used, together with the behavioural thresholds $\vartheta$ following inactivations of different brain areas, to infer the relative scaling of their weights $\boldsymbol{a}$. First we describe the main approach in the general setting with multiple populations, and then we specialize to the particular case of two populations and apply it to our data.

Previous work has shown how one can combine knowledge of choice correlations and neural noise correlations to estimate the decoding weights of individual neurons[14]. If decoded neural responses in each population are dominated by a single mode, then we can extend this concept to the population level. The population-level analog of a neural response $r_k$ is an estimate $\hat{s}_x$ derived from population $x$. The analog of choice correlations $C_k$ are the slopes $\beta_x$ that relate observed and optimal choice correlations, and the analog of noise covariance $\Sigma_{ij}$ between neurons $i$ and $j$ is the covariance $\varepsilon_{xy}$ (**Eqs (11) & (14)**) between estimates $\hat{s}_x$ and $\hat{s}_y$ derived from distinct populations.

Unlike neural noise correlations, we cannot directly measure the noise correlations $E$ at the population level. Nonetheless, we can infer those population-level noise correlations indirectly from inactivation experiments, in which behavioral thresholds are measured after altering the decoder scaling afforded to different brain areas by a factor $\rho_{x\phi}$ for inactivation experiment number $\phi$. In our feedforward linear model, it is mathematically equivalent to reduce the activity by $\rho_{x\phi}$, or to alter a decoder's scaling $a_x$ by the same factor. Totally inactivating an area is equivalent to setting its scaling to zero, but here we permit partial inactivation of multiple brain areas. For now, we assume these inactivation factors are controlled by the experimenter, and thus known, although later we will incorporate some uncertainty about these inactivations.

Each such experiment provides one constraint on the unknown population properties, according to

$$\theta_\phi^2 \approx \frac{\boldsymbol{a}_\phi \cdot E \cdot \boldsymbol{a}_\phi}{|\boldsymbol{a}_\phi|_{l_1}^2} = \frac{1}{\left(\sum_x a_x\rho_{x\phi}\right)^2}\sum_{xy} a_x\rho_{x\phi}E_{xy}\rho_{y\phi}a_y \tag{18}$$

where $\theta_\phi$ is the behavioural threshold during the $\phi$'th inactivation experiment, $\boldsymbol{a}_\phi$ is the vector of decoder scaling factors for the different populations with components $a_{x\phi} = a_x\rho_{x\phi}$, and where the $l_1$-normalization $|\boldsymbol{a}_\phi|_{l_1} = \sum_x a_x\rho_{x\phi}$ ensures that the decoder remains unbiased after inactivation (as observed experimentally[18,22]). In such experiments one could also measure the slopes $\beta_{x\phi}$ of the choice correlations for multiple different populations to provide additional measurement constraints

$$\beta_{x\phi} \approx \frac{\delta_x \cdot E \cdot \boldsymbol{a}_\phi}{|\boldsymbol{a}_\phi|_{l_1}} = \frac{1}{\sum_x a_x\rho_{x\phi}} \sum_y E_{xy}\rho_{y\phi}a_y \tag{19}$$

Notice that **Eqs (18)** and **(19)** can be written as multivariate polynomials up to cubic order jointly in the unknowns $E$ and $\boldsymbol{a}$. Altogether there are $Z(Z+1)/2$ unknowns for the covariance matrix $E$, and another $Z$ unknowns for the intact brain's decoder scaling factors $\boldsymbol{a}$. As long as the number of independent threshold and slope measurements is at least as large as the number of unknowns, then **Eq (19)** can be solved numerically (S2 Fig), revealing the correct decoder scaling for multiple populations. Slopes of choice correlations during inactivation experiments provides a larger number of data points from a given set of inactivation experiments than measuring the thresholds alone.

**Two population solution.** When only two populations of neurons, $x$ and $y$, are relevant for a particular task, this general approach to identifying their relative scaling can be simplified. We next describe this simpler two-population theory, and then apply it to data from the vestibular system.

If we can completely inactivate one brain area, then from **Eq (1)**, the animal's total estimate $\hat{s}$ would be equal to either $\hat{s}_x$ or $\hat{s}_y$, depending on which area is inactivated. The resultant behavioural threshold would simply reflect the variance of the remaining estimate, which is equal to the magnitude of dominant decoded noise within the active area, so $\vartheta^2_{-x} \approx \varepsilon_{yy}$ and $\vartheta^2_{-y} \approx \varepsilon_{xx}$. If populations $x$ and $y$ are uncorrelated ($\varepsilon_{xy} = 0$), then the ratio of weight scaling factors can be factorized into a product of ratios (S5 Text):

$$\frac{a_x}{a_y} = \frac{\beta_x}{\beta_y}\frac{\varepsilon_{yy}}{\varepsilon_{xx}} \approx \frac{\beta_x}{\beta_y}\frac{\vartheta^2_{-x}}{\vartheta^2_{-y}} \tag{20}$$

where the two independent factors represent outcomes of correlational and causal studies. If readout is optimal, then the multipliers $\beta_x$ and $\beta_y$ are both equal to one, so $a_x/a_y = \vartheta^2_{-x}/\vartheta^2_{-y}$. This is consistent with the general belief that the behavioural effects of inactivating a brain area must be commensurate with its contribution to the behaviour. A departure from optimality could break this relationship, so the effects of causal manipulation may not match the relative sensitivities of the brain areas (S3 Fig). Even in purely feedforward networks, the magnitude of neuronal choice correlations need not equal the effects of inactivation. Thus, disagreements between the two experimental outcomes should not be entirely surprising and do not undermine the functional significance of either.

In fact, **Eq (20)** revealed how one can combine choice correlations and behavioural thresholds to infer the contributions of two uncorrelated areas. But if the areas are correlated, one must explicitly account for the magnitude of correlation between areas $\varepsilon_{xy}$ and the ratio of scales no longer factorizes:

$$\frac{a_x}{a_y} \approx \left(\frac{\beta_x}{\beta_y}\frac{\vartheta^2_{-x}}{\vartheta^2_{-y}} - \gamma\right)\left(1 - \frac{\beta_x}{\beta_y}\gamma\right)^{-1} \tag{21}$$

where $\gamma = \varepsilon_{xy} / \varepsilon_{xx}$ is the magnitude of correlated noise between the two populations' estimates relative to the variance of estimates from $x$ alone. Note that one can also use **Eqs (20)** and **(21)** to compute the optimal weight scaling factors simply by setting both $\beta_x$ and $\beta_y$ to 1. Therefore, we can use these equations not only to determine the relative weights of brain areas but to also to evaluate precisely how suboptimal those weights are.

## Application to data

We now use the techniques developed so far to infer the relative contributions of two brain areas in macaque monkeys to heading discrimination. Data were collected from monkeys trained to discriminate their direction of self-motion in the horizontal plane (**Fig 2A**) using vestibular (inertial motion) and/or visual (optic flow) cues (see **Methods**; see also refs. [21,23]). At the end of each trial, the animal reported whether their perceived heading $\hat{s}$ was leftward ($\hat{s} < 0^{\circ}$) or rightward ($\hat{s} > 0^{\circ}$) relative to straight ahead.

**Discrepancy between correlation and causal studies.** Responses of single neurons were recorded from either area MSTd (monkeys A and C; $n$=129) or area VIP (monkeys C and U; $n$=88) during the heading discrimination task (see Methods). Basic aspects of these responses were analyzed and reported in earlier work[21,23]. Briefly, it was found that neurons in VIP had substantially greater choice correlations (CC) than those in MSTd (Fig 2B – left) for both the vestibular and visual conditions. This difference in CC between areas could not be attributed to differences in neuronal thresholds $\vartheta_k$ (Fig 2B – middle), defined as the stimulus magnitude that can be discriminated correctly 68% of the time ($d'$=1) from neuron $k$'s response $r_k$ (Methods; S3 Fig). Based on its greater CCs, one might expect that VIP plays a more important role in heading discrimination than MSTd. In striking contrast to this expectation, a recent study showed that there was no significant change in heading thresholds following VIP inactivation for either the visual or vestibular stimulus conditions[18] (Fig 2B – right (blue); monkeys B and J). On the other hand, inactivation of MSTd using a nearly identical experimental protocol led to substantial deficits in heading discrimination performance[22] (Fig 2B – right (red); monkeys C, J, and S). The neural and inactivation studies in VIP used non-overlapping subject pools, so the observed dissociation between CCs and inactivation effects could potentially reflect the idiosyncrasies of the subjects' brains. To rule this out, we repeated the inactivation experiment by specifically targeting Muscimol injections to sites in area VIP that were previously found to contain neurons with high CCs in another monkey and obtained similar results (S5 Fig).

These findings reveal a striking dissociation between choice correlations and effects of causal manipulation: VIP has much greater CCs than MSTd yet inactivating VIP does not impair performance. One may be tempted to simply conclude that VIP does not contribute to heading perception. We will now show that this is not necessarily true. Depending on the structure of correlated noise and the decoding strategy, neurons in both areas may be read out in a manner that is entirely consistent with the observed effects of inactivation.

**Test for optimality.** We first asked if the above results can simply be explained if the brain allocated weights optimally to the two areas. To answer this, we tested if neuronal choice correlations satisfied Eq (4). Binary discrimination experiments typically do not measure choice correlations $C_k = \mathrm{Corr}(r_k, \hat{s} | s = s_0)$ because they do not have direct access to the animal's continuous stimulus estimate $\hat{\boldsymbol{s}}$; they only track the animal's binary choice. Instead they measure a related quantity known as choice probability defined as the probability that a rightward choice is associated with an increase in response of neuron $k$ according to $CP_k = P(r_k^+ > r_k^-)$ where $r_k^{\pm} \sim P(r_k | \mathrm{sgn}(\hat{s}) = \pm 1)$ is a response $r_k^{\pm}$ of neuron $k$ when the animal chooses ±1. Therefore we first transformed the measured choice

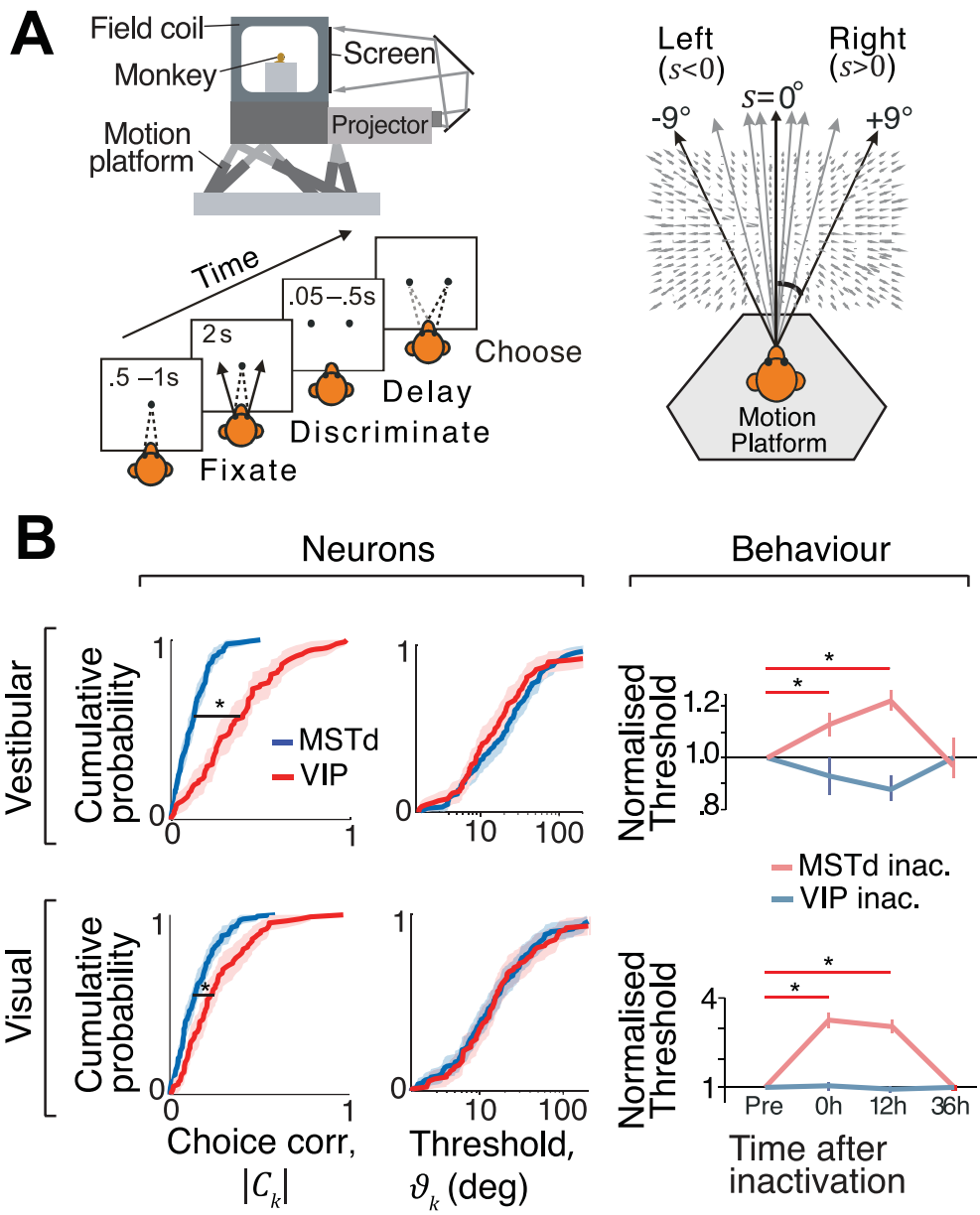


**Fig 2. Choice-related activity and effects of inactivation. (A)** Behavioural task: the monkey sits on a motion platform facing a screen. He fixates on a small target at the center of the screen, and then we induce a self-motion percept by moving the platform (vestibular condition) or by displaying an optic flow pattern on the screen (visual condition). The fixation target then disappears and the monkey reports his percept by making a saccade to one of two choice targets. **(B) Left**: Neurons in both MSTd ($n$=129) and VIP ($n$=88) exhibited significant choice correlations (CCs). The median CC of VIP neurons was significantly greater than that of MSTd neurons ($^*p<0.001$, Wilcoxon rank-sum test) in both vestibular (top) and visual (bottom) conditions. **Middle**: Median neuronal thresholds were not significantly different between areas (vestibular: $p$=0.94, visual: $p$=0.86, Wilcoxon rank–sum test). **Right**: Average discrimination thresholds at different times relative to inactivation of VIP (unsaturated blue) and MSTd (unsaturated red). All threshold values were normalized by the corresponding baseline thresholds ("pre"). Shaded regions and error bars denote standard errors of the mean (SEM); asterisks indicate significant differences ($^*p<.05$, $t$–test). Neural data re-analyzed from refs. [21,23]. Inactivation data reproduced from refs. [18,22].

https://doi.org/10.1371/journal.pcbi.1006371.g002

probabilities to choice correlations using a known relation[14] before further analyses (Methods). Equivalently, one could measure the correlation $\mathrm{Corr}(r_k, \mathrm{sgn}(\hat{s})|s = s_0)$ between the neural response and the binary choice, which [15] showed is $\approx 0.8C_k$. Note

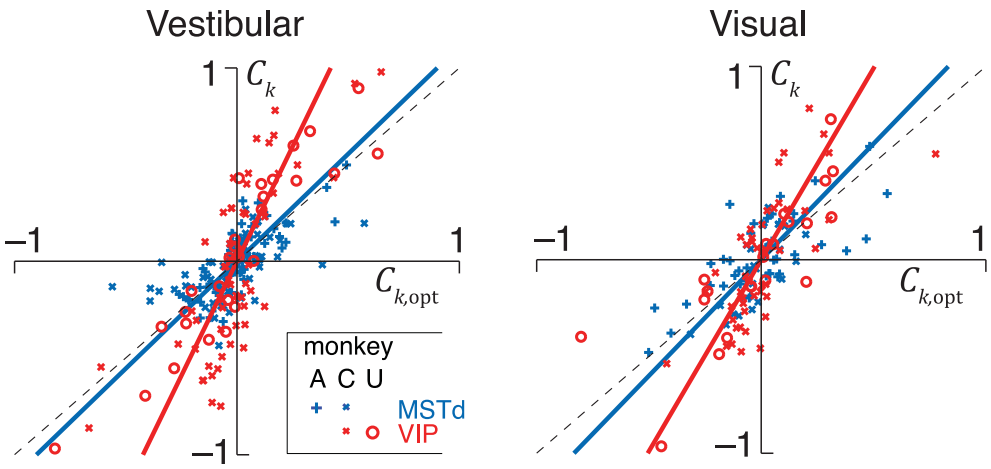


**Fig 3. Readout is not optimal.** Whereas the experimentally measured choice correlations ($C_k$) of neurons in MSTd (blue) for both the vestibular (left) and the visual (right) condition are well described by the optimal predictions ($C_{k,\mathrm{opt}}$), those of VIP neurons are systematically greater (red). This observation was consistent across all monkeys (see S5A Fig for monkey *X*). Solid lines correspond to the best linear fit. Vestibular data replotted from Ref.[15] with different sign convention (see Methods).

https://doi.org/10.1371/journal.pcbi.1006371.g003

that the above definition gives choice correlations that are either positive or negative depending on whether a rightward choice is associated with an increase or decrease in neuronal response. Therefore, we adjusted Eq (4) to generate predictions for optimal CCs that accounted for our convention (see **Methods**).

Fig 3 compares experimentally measured CCs against the CCs predicted by optimal decoding for all neurons recorded in the vestibular (left panel) and visual (right panel) conditions (see S6 Fig for data from individual animals). Our data are consistent with optimal decoding of MSTd, since the predicted and measured CCs are significantly correlated (vestibular: Pearson's $r$ =0.65, $p<10^{-3}$; visual: $r$ =0.70, $p<10^{-3}$) with a slope not significantly different from 1 (vestibular: slope = 1.11, 95% confidence interval (CI) =[0.83 1.54]; visual: slope = 1.24, 95% CI =[0.94 1.78]). For VIP, although the predicted and measured CCs are again strongly correlated (vestibular: $r$ = 0.80, $p<10^{-3}$; visual: $r$ = 0.75, $p<10^{-3}$), the regression slope deviates substantially from unity (vestibular: slope=2.37, 95% CI =[1.97 3.08]; visual: slope=1.98, 95% CI =[1.41 2.74]), demonstrating that our data are inconsistent with optimal decoding. Note that, if VIP is decoded suboptimally, this implies that the overall decoding—one based on both VIP and MSTd—is suboptimal as well because the decoder failed to use all information available in the neurons across both populations. This leads to two questions: First, how much information is lost by suboptimal decoding? Second, how is this information lost? To get precise answers, we will now determine how the brain weights activity in MSTd and VIP to perform heading discrimination.

**Inferring readout weights.** Throughout this section, we use subscripts *M* and *V* to denote MSTd and VIP instead of the generic subscripts *x* and *y* used to describe the methods. For clarity, we will restrict our focus to the vestibular condition but results for the visual condition are presented in the supporting information. In order to determine decoding weights, we constructed two kinds of covariance structures that implied either extensive or limited information as explained earlier.

In the extensive information case, we modeled noise covariance using data from pairwise recordings within MSTd and VIP reported previously [21,29]. Those experiments established that noise correlation between neurons in these areas tends to increase linearly with the

similarity of their tuning functions, or signal correlation (**Eq (8)**). This relationship between noise and signal correlations has a substantially steeper slope in VIP than in MSTd (MSTd: $m_M = 0.19 \pm 0.08$; VIP: $m_V = 0.70 \pm 0.16$, S7 Fig). We used these empirical relationships to extrapolate noise correlations between all pairs of independently recorded neurons within each of the two populations, using only their tuning curves, and assuming that any stimulus-dependent changes in correlation were negligible. Although the neural sensitivities were comparable in the two brain areas, the stronger correlations in VIP gave it higher information content than MSTd: since the dominant noise modes point away from the signal direction, greater correlations lead to less noise variance along the signal direction, and hence more information [35]. Since correlations between VIP and MSTd populations were not measured experimentally, we explored different correlation matrices (see **Methods**, **Eq (24)**).

In the limited information case, we added correlations that limited the total information content across the two populations (**Eq (13)**). For this latter case, we relied on behavioural thresholds before and after inactivation, and choice correlations, to determine the magnitudes of noise within ($\varepsilon_{MM}$ and $\varepsilon_{VV}$) and between ($\varepsilon_{MV}$) areas (see **Methods**). In both cases, we constructed covariances for many different population sizes $N$ by sampling equal numbers of neurons from both areas with replacement. The choice of distributing neurons equally among the two areas was made only for convenience and has no bearing on the result as explained later.

Fig 4A shows example covariance matrices for both extensive and limited information models for a population of 128 neurons. The two structures look visually similar because the additional fluctuations caused by information-limiting correlations are quite subtle. Nevertheless, there is a huge difference between the two models in terms of their information content (Fig 4B). The extensive model has information that grows linearly with $N$, implying that these brain areas have enough information to support behavioural thresholds that are orders of magnitude better than what is typically observed. However, when information-limiting correlations are added, information saturates rapidly suggesting that behavioural thresholds may not be much lower than population thresholds even if the decoding weights are fine-tuned for best performance. We will now infer scaling factors $a_M$ and $a_V$ of decoding weights using both noise models and examine their implications.

**Extensive information model.** We've already seen that the pattern of choice correlations is not consistent with optimal decoding of MSTd and VIP. In fact, for the extensive information model, optimal decoding will lead to extremely small CCs by suppressing response components that lie along the leading noise modes as they have very little information (S8A Fig). Ironically, the magnitude of CCs found in our data could only have emerged if the response fluctuations along those leading modes substantially influenced animal's choice (S8B Fig). This means that the decoder must be largely confined to the subspace spanned by those modes. We therefore restricted our focus to the two leading eigenvectors $\mathbf{u}^1$ and $\mathbf{u}^2$ of the covariance matrix. When the two populations are uncorrelated, these vectors lie exclusively within the one-dimensional subspaces spanned by neurons in MSTd and VIP respectively (Fig 5A). In our case, vectors $\mathbf{u}^1$ and $\mathbf{u}^2$ corresponded to $\mathbf{u}^V$ and $\mathbf{u}^M$. Although decoding only this subspace is not optimal with respect to the total information content in the two areas, a decoder could still be optimal within that subspace. To test this, we estimated the choice correlations $C^V_{k,\mathrm{opt}}$ and $C^M_{k,\mathrm{opt}}$ that would be expected from optimally weighting the two areas within this subspace (Eq (7)). The observed CCs were proportional (MSTd: Pearson's $r = 0.55$, $p < 10^{-3}$; VIP: $r = 0.76$, $p < 10^{-3}$) to these optimal predictions implying that the leading noise modes of the extensive information model are able to capture the basic structure of choice-related activity in both areas (Fig 5B). However the slopes $\beta_M$ and $\beta_V$ were significantly different from 1 ($\beta_M = 0.73$, 95% CI =[0.63 0.84]; $\beta_V = 2.38$, 95% CI = [2.2 2.57]) implying that the weight scaling factors $a_M$ and $a_V$ must be suboptimal

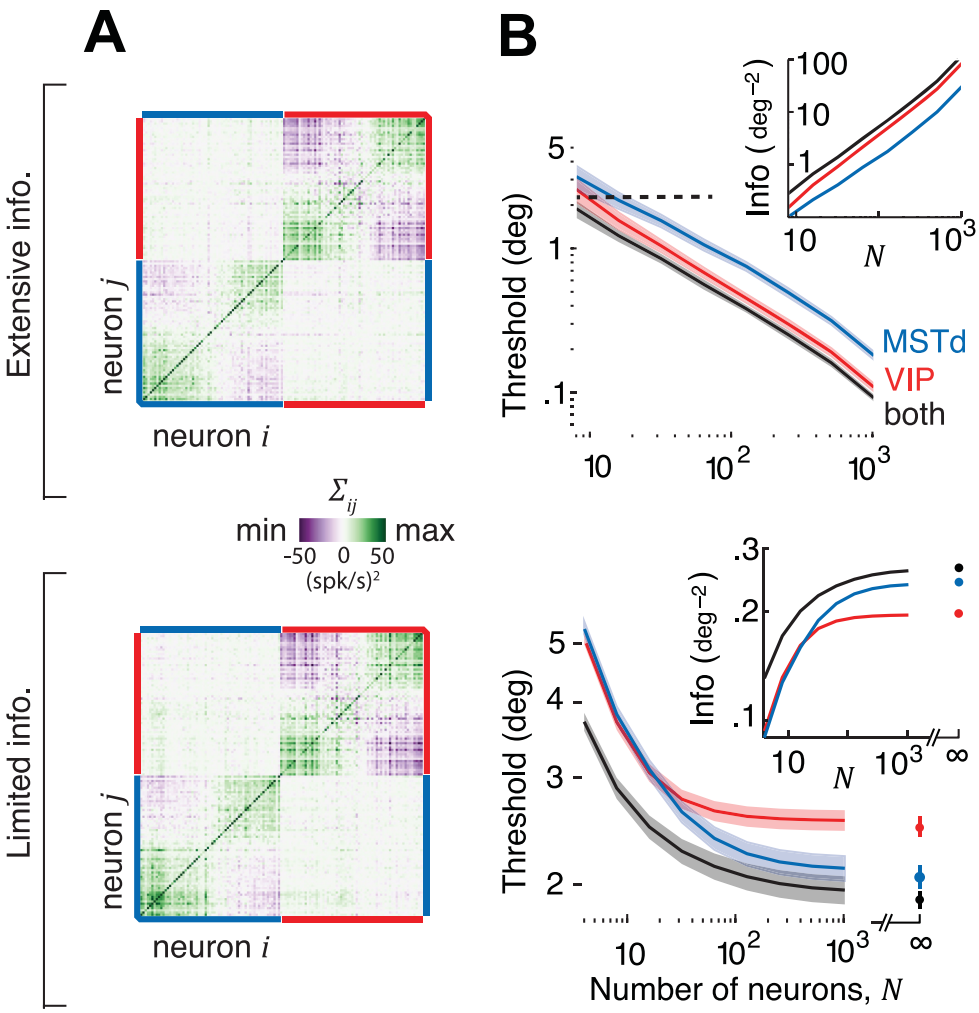


**Fig 4. Covariance structure of extensive and limited information models. (A)** Matrix of covariances $\Sigma_{ij}$ among neurons in MSTd and VIP ($N$=128). Top: Extensive information model constructed by sampling according to the empirical relationship in S7 Fig, for the case when the two areas are uncorrelated on average. Bottom: Limited information model adds a small amount of information-limiting correlations with magnitudes ($\varepsilon_{MM} = 4.2$, $\varepsilon_{VV} = 7$, $\varepsilon_{MV} = 0$) chosen arbitrarily for illustration. **(B)** Inset shows the effect of population size on the information content implied by the two kinds of noise in MSTd (blue), VIP (red) and in both areas together (black). If decoded optimally, behavioural thresholds implied by the extensive information model would decrease with $N$ resulting in performance levels that are vastly superior to those actually observed in monkeys (black dashed line). Information-limiting correlations cause information to saturate with $N$.

https://doi.org/10.1371/journal.pcbi.1006371.g004

even within the two-dimensional subspace. Since we knew the magnitudes of $\varepsilon_{MM}$ and $\varepsilon_{VV}$ for this noise model from pairwise recordings (Table 1), we applied the exact rather than approximate form of Eq (20) and obtained a scaling ratio $a_M/a_V = 0.8 \pm 0.1$.

To test whether the inferred scaling was meaningful, we compared behavioural thresholds implied by the resulting decoding scheme against experimental findings of inactivation. The threshold prior to inactivation is related to the variance of the estimator whose decoding weights $\mathbf{w}$ are along the direction specified by $a_M\mathbf{u}^M + a_V\mathbf{u}^V$. Inactivating either area is equivalent to setting the corresponding scaling factors to zero, so post-inactivation thresholds are given by the variance along the leading noise mode specific to the active area ($\mathbf{u}^M$ or $\mathbf{u}^V$). We computed pre and post-inactivation thresholds and found they were qualitatively consistent with experimental results: for large populations, MSTd inactivation is predicted to produce a

**Table 1. Model parameters and predicted changes in CCs following inactivation for the two covariance models, shown as median ± central quartile range.** (†Values correspond to when decoder is inferred using a rank-two approximation of the covariance.).

| Model | | Extensive information model† | Limited information model |
|---|---|---|---|
| **Model parameters** | Noise magnitudes | $\varepsilon_{MM} = 15, \varepsilon_{VV} = 45, \varepsilon_{MV} = 0$ | $\varepsilon_{MM} = 5, \varepsilon_{VV} = 38, \varepsilon_{MV} = 10$ |
| | Multiplicative scaling of CCs relative to optimal | $\beta_M = 0.44, \beta_V = 1.4$ | $\beta_M = 1.1, \beta_V = 2.4$ |
| | Optimal weights | $\lvert a_M/a_V \rvert = 2.8 \pm 0.5$ | $\lvert a_M/a_V \rvert = 9 \pm 4$ |
| | Inferred weights | $\lvert a_M/a_V \rvert = 0.8 \pm 0.1$ | $\lvert a_M/a_V \rvert = 14 \pm 7$ |
| **Model predictions** | Multiplicative change in CCs following inactivation | $\zeta_M = 2.2 \pm 0.3$ | $\zeta_M = 0.9 \pm 0.4$ |
| | | $\zeta_V = 1.3 \pm 0.1$ | $\zeta_V = 1.3 \pm 0.4$ |

https://doi.org/10.1371/journal.pcbi.1006371.t001

large increase in threshold (Fig 5C, red vs black) whereas VIP inactivation is predicted to have little or no effect (Fig 5C, blue vs black; see S9 Fig for visual condition). This correspondence

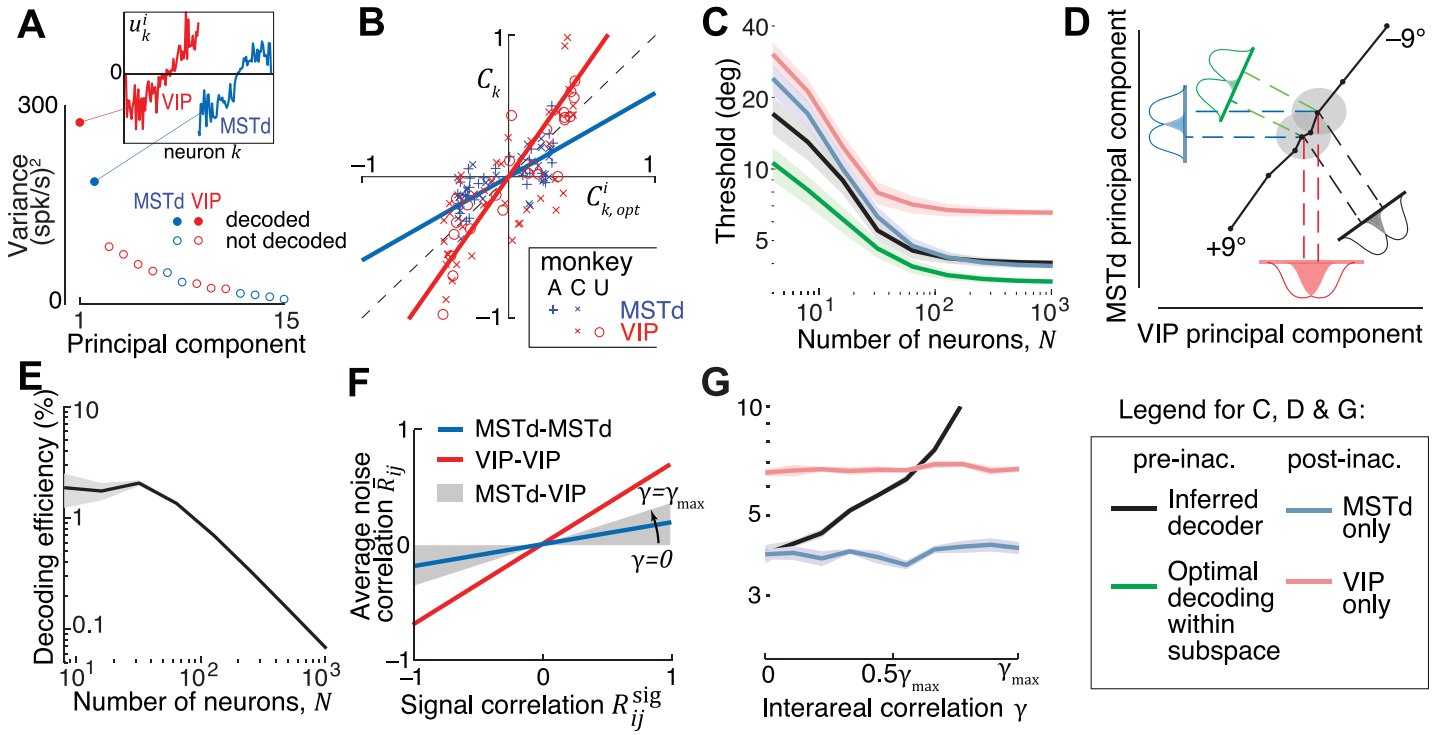


**Fig 5. Decoder inferred using the extensive information model. (A)** Decoding weights were inferred in the subspace of 2 leading principal components of noise covariance (solid circles). Inset: These components lie entirely within the space spanned by neurons in one of the two brain regions. Components are color coded according to the brain region that it inhabits (red=VIP; blue=MSTd). **(B)** Experimentally measured choice correlations ($C_k$) of individual neurons in VIP (red) and MSTd (blue) are plotted against their respective components $C^1_{k,\mathrm{opt}}$ and $C^2_{k,\mathrm{opt}}$ of choice correlations generated from optimally decoding responses within the subspace of 2 leading principal components. **(C)** Unlike the optimal decoder in Fig 4B, the behavioural threshold predicted by the inferred weights (black) saturates at a population size of about 100 neurons. The green line indicates the performance of an optimal decoder within the two-dimensional subspace. Inactivating VIP is correctly predicted to have no effect on behavioural performance for large $N$ (blue), while MSTd inactivation increases the threshold (red). **(D)** A schematic of the inferred decoding solution projected onto the first principal component of noise in VIP and MSTd. The solid colored lines correspond to the readout directions for the four cases shown in (c). The long diagonal black line is the projection of the mean population responses for headings from –9˚ to +9˚, and the two gray ellipses correspond to the noise distribution at heading directions of ±2˚. The colored gaussians correspond to the projections of this signal and noise onto each of the four readout directions, and the overlap between these gaussians corresponds to the probability of discrimination errors. **(E)** The percentage of available information read out by the inferred decoder (the decoding efficiency) decreases with population size, because the decoded information saturates while the total information is extensive. **(F)** Correlations between MSTd and VIP were not measured experimentally. We modeled these correlations according to the same linear trend that on average described correlations within each population, but with different slopes, yielding different interareal correlations parametrized by $\gamma = \varepsilon_{MV}/\varepsilon_{MM}$ (**Methods**). This slope reaches its maximum allowable value $\gamma_{\max} = \sqrt{\varepsilon_{VV}/\varepsilon_{MM}}$, the geometric mean of the slopes for MSTd and VIP. **(G)** For each value of $\gamma$, we used the resultant covariance and CCs to infer the decoder, and plotted its behavioural thresholds. Thresholds are shown for a population of 256 neurons, by which point the performance had saturated to its asymptotic value for all $\gamma$. Shaded regions in (c), (e), and (g) represent ±1 SEM.

https://doi.org/10.1371/journal.pcbi.1006371.g005

to experimental inactivation results is remarkable because the procedure to deduce scaling factors $a_M$ and $a_V$ was not constrained in any way by behavioural data, but rather informed entirely by neuronal measurements. We also confirmed that the threshold expected from optimal scaling factors (Table 1) was smaller than that produced by inferred weights (Fig 5C, green vs black) implying that the brain indeed weighted the two areas suboptimally.

The above findings are explained graphically in Fig 5D by projecting the relevant quantities (tuning curves $\mathbf{f}(s)$, noise covariance $\Sigma$, decoding weights $\mathbf{w}$) onto the subspace of the first two principal components ($\mathbf{u}^M$ and $\mathbf{u}^V$) of the noise covariance $\Sigma$. The colored lines indicate different readout directions, determined by the scaling ($a_M$ and $a_V$) of weights for the two populations. A ratio of $|a_M/a_V| > 1$ corresponds to greater weight on the estimate derived from MSTd activity, and the associated readout direction will be closer to the principal component of MSTd. The response distributions are depicted as gray ellipses (isoprobability contours) for the two stimuli to be discriminated. The discrimination threshold for different decoders can be obtained simply by projecting these ellipses onto the readout direction of the specified decoder and examining the overlap between the projections. Within this subspace, the ratio $|a_M/a_V|$ of the decoder inferred from CCs was much smaller than the optimal ratio (Table 1), meaning that MSTd was given too little weight. Consequently, the response distributions have more overlap along the direction corresponding to the decoder inferred from neuronal CCs (black) than along the optimal direction in that subspace (green). This means that the outputs are less discriminable and thus that the decoding is suboptimal. VIP inactivation ($a_V = 0$) corresponds to decoding only from MSTd (blue). This happens to produce no deficit because the overlap of the response distributions is similar to that along the original decoder direction. On the other hand, inactivating MSTd ($a_M = 0$) corresponds to decoding only from VIP (red), where the two response distributions have greater overlap leading to a larger threshold.

It is important to keep in mind that decoding the noisiest two-dimensional subspace, which throws away all signal components in the remaining low-noise $N$–2 response dimensions, is a much more severe suboptimality than misweighting the two areas' signals within that restricted subspace, which loses less than half the information (Fig 5C). As illustrated in Fig 5E, the efficiency — the fraction of available linear Fisher information recovered by this decoder ($\eta = J_{\text{decoded}}/J_{\text{opt}}$) — drops precipitously with the number of neurons ($\eta \sim 2.5N^{-1}$). Moreover, for this model, a steeper relationship between signal and noise correlations leads to greater CCs. This is because the model is only consistent with suboptimal decoding that fails to remove the strong noise correlations; these noise correlations are decoded to drive the choice, and thus correlate neurons not only with each other but also with that choice. Thus, in the extensive information model, high CCs are a consequence of decoding a restricted subspace of neural activity, a radically suboptimal strategy for the brain.

Behavioural predictions of this model were robust to assumptions about the exact size of the decoded subspace (S10 Fig), but were found to depend on the magnitude of noise correlations between the VIP and MSTd populations. Since interareal correlations were not measured, we systematically varied the strength of these correlations by changing $\gamma$ (Fig 5F), and used **Eq (21)** to infer scaling factors for each case. We used these scaling factors to generate behavioural predictions for different values of $\gamma$. Predictions for one example value of these correlations are shown in S11 Fig. Behavioural predictions progressively worsened as a function of the strength of noise correlations between MSTd and VIP: for this model, even weak but nonzero interareal correlations imply that inactivating area VIP should improve behavioural performance (Fig 5G).

**Limited information model.** In the presence of information-limiting correlations, choice correlations must be proportional to the ratio of behavioural to neuronal thresholds (Eq (17)). This was indeed the case both in MSTd and VIP as we showed already in Fig 3. Those slopes

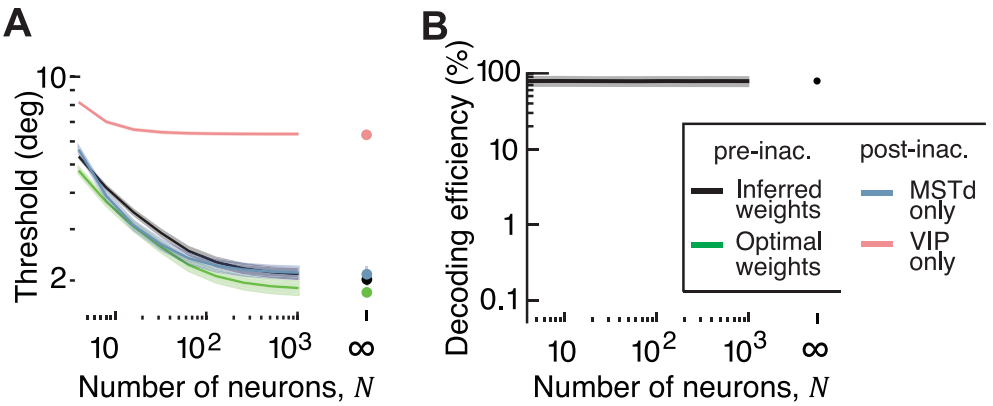


**Fig 6. Decoder inferred using the limited information model. (A)** Like decoding in the presence of extensive information, this decoder is suboptimal (black vs green), and can account for the behavioural effects of inactivation. **(B)** Unlike decoding in the extensive information model, the efficiency of this decoder (expressed in percentage) is high and insensitive to population size. Shaded areas represent ±1 SEM.

https://doi.org/10.1371/journal.pcbi.1006371.g006

correspond to the multipliers $\beta_M$ and $\beta_V$ for this model, and were found to be different for the two areas (Table 1).

As we noted earlier, unlike the leading modes of noise in the extensive information model, the magnitudes of information-limiting correlations ($\varepsilon_{MM}$, $\varepsilon_{VV}$ and $\varepsilon_{MV}$) are difficult to measure. Nevertheless, we can deduce them from behaviour because behavioural precision is ultimately limited by these correlations. Briefly, using behavioural thresholds *after* inactivation of each area, along with $\beta_M$ and $\beta_V$ derived from choice correlations as additional constraints, we can simultaneously infer the magnitude of information-limiting correlation within each area ($\varepsilon_{MM}$ and $\varepsilon_{VV}$), the correlated component of the noise ($\varepsilon_{MV}$), and scaling factors ($a_M$ and $a_V$) (see **Methods**). A model based on these inferred parameters correctly predicted that the behavioural threshold *before* inactivation would not be significantly different from threshold following VIP inactivation (**Fig 6A**; see **S12 Fig** for visual condition). This was because the scaling of weights in MSTd was much larger than in VIP according to this model ($a_M \gg a_V$, **Table 1**), so inactivating VIP had little impact on the output of the decoder and left behaviour nearly unaffected. Unlike the decoder inferred for the extensive information model, the efficiency $\eta$ of this decoder did not depend on the size of the population being decoded (**Fig 6B**, $\eta = J_{\text{decoded}}/J_{\text{opt}} = \theta^2_{\text{opt}}/\theta^2_{\text{decoded}} = (1.98 \pm 0.06)^2/(2.2 \pm 0.17)^2 = 0.79 \pm 0.13$) because neurons in this model carry a lot of redundant information.

**Effect of temporal variability.** All analyses above were performed on neural data in the central 400ms of the trials following earlier work. This corresponds to an implicit assumption that monkeys made their decisions based solely on the information available during the period of the trial where the stimulus amplitude was highest (Gaussian stimulus profile). However, the experiments did not measure the monkeys' psychophysical kernel, so we do not know if the above assumption is strictly valid. Moreover, both stimulus and choice-related activity typically vary across time in MSTd [23] and VIP [21], so it is unclear if our conclusions about the relative decoding weights hold outside of the time-window considered in the above analysis. To test this, we repeated our analysis using a sliding window to estimate decoding weights across time. As expected, both neuronal thresholds (Fig 7A) and choice correlations (Fig 7B) were variable across time. Transiently higher firing rates at stimulus onset provide more information early in a trial, but choice correlations peak in the middle of the stimulus. Consequently, the slopes relating observed and optimal choice correlations also varied over time in

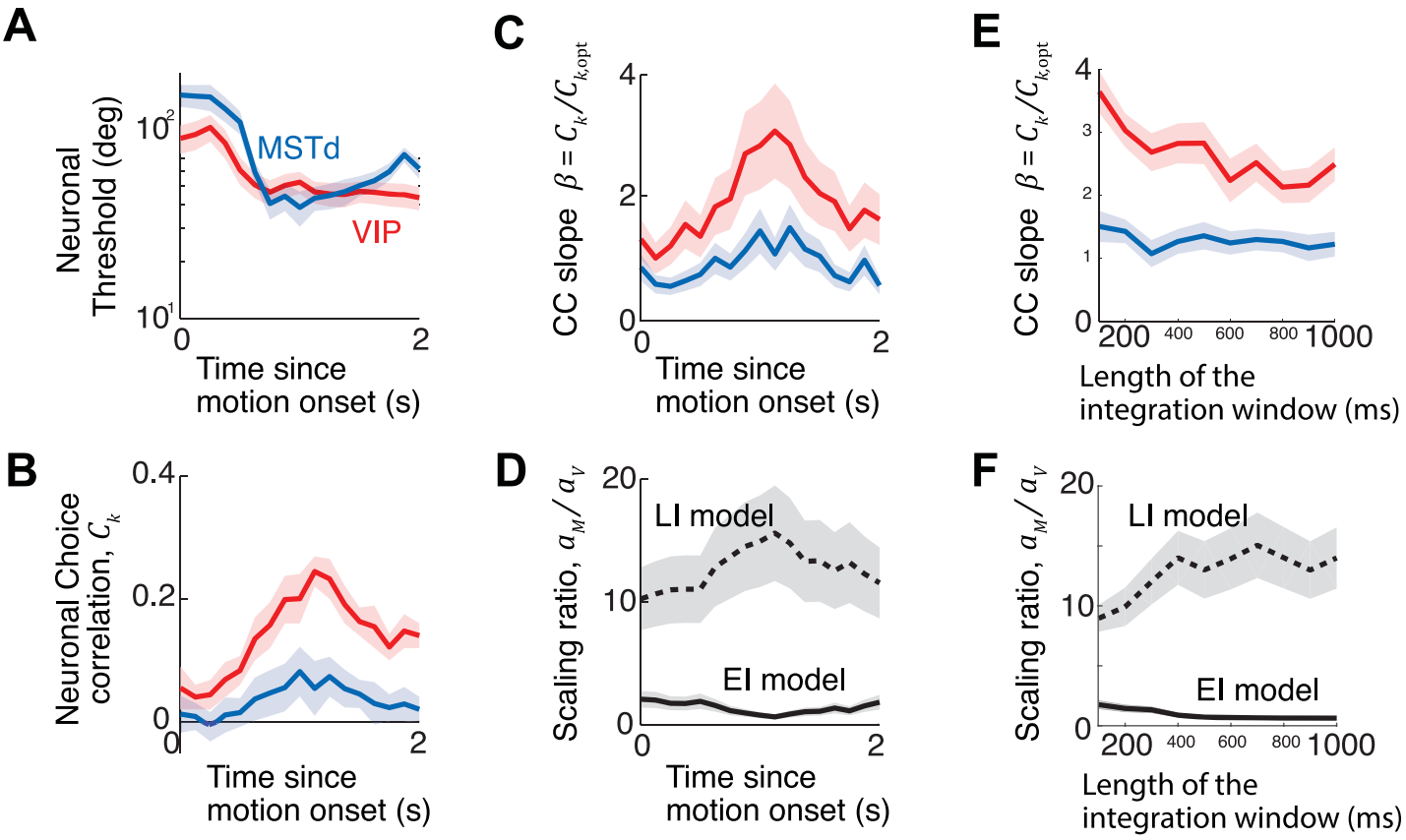


**Fig 7. Readout weights do not vary drastically across time.** Neuronal thresholds **(A)** and choice correlations **(B)** were computed for each neuron across the duration of the trial using a 250ms moving window and averaged across neurons. Note that these readouts predict the choice based only on a single time window per data point, and do not perform a weighted sum of responses in multiple windows. Neuronal thresholds in both brain areas were comparable at all times, yet the choice correlations (CCs) differed between brain areas VIP and MSTd in a consistent manner over time. Although CCs in both areas peaked around the middle of the trial, those in VIP were proportionally larger at almost all times. **(C)** Consequently the slopes, $\beta = C_k/C_{k,opt}$, that related observed and optimal choice correlations were generally greater in area VIP than in MSTd. **(D)** The readout weights inferred using the two models remain largely constant throughout the trial, and are qualitatively consistent with the conclusions drawn from our analyses presented in the main text: the extensive information model implies that area MSTd is underweighted, whereas the limited information model predicts the opposite. Symbols $a_M$ and $a_V$ denote scaling of readout weights of areas MSTd and VIP respectively. **(E)** Regression slopes are minimally affected by the length of the analysis window. Both observed neuronal choice correlations as well as those implied by optimal decoding of MSTd and VIP populations increased similarly with the length of the analysis window. This leaves the regression slopes $\beta = C_k/C_{k,opt}$ largely invariant with the window length for both VIP (red) and MSTd (blue). **(F)** The qualitative difference in the readout weights inferred using the two noise models are consistent across different lengths of analysis window. Error bars denote ±1 standard deviation. See **S13 Fig** for visual condition.

https://doi.org/10.1371/journal.pcbi.1006371.g007

both areas (Fig 7C). Nevertheless, the time-course of the ratio of scaling factors was much less variable and the qualitative differences in the extensive and limited information models described above are still found to hold throughout the trial (Fig 7D). A full model of the time course of these signals will likely require recurrence for temporal integration (see Discussion). However, temporal integration of independent evidence would yield choice correlations that should grow monotonically with time, so the observed dynamics already indicate another form of suboptimality. Decoding weights may also depend on the length of the integration window and past studies have proposed ways to simultaneously infer the length of integration window and decoding weights from neural data [32]. Although we did not infer the size of the integration window, we found that the slopes of choice correlations in VIP were larger than MST for various choices of integration window, implying that our conclusions are robust to the duration of the analysis window (Fig 7E).

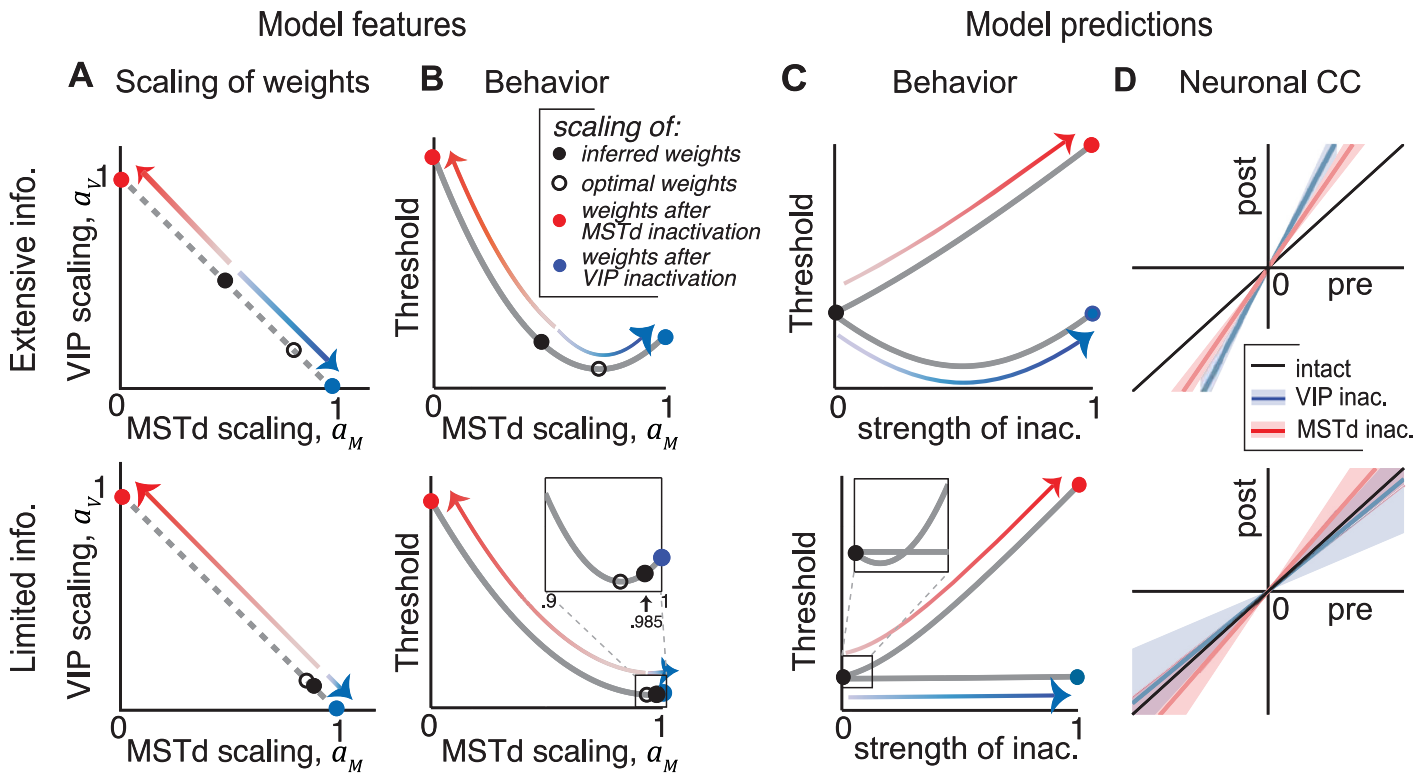


**Fig 8. Decoding strategy and model predictions for the extensive information model and the limited information model. (A)** Optimal (open black) and inferred (filled black) scaling of weights in MSTd ($a_M$) and VIP ($a_V$). Inactivation of either MSTd (red) or VIP (blue) confines the readout to the active area resulting in a scaling of 1. Red and blue arrows indicate the transformation resulting from inactivating MSTd and VIP respectively. The scaling factors always sum to 1. **(B)** Behavioural threshold $\vartheta$ as a function of $a_M$. Whereas $\vartheta$ increases following MSTd inactivation for both models (red), it improves initially following partial VIP inactivation (blue) in the extensive information model (top) but remains unchanged in the limited information model (bottom). **(C)** The same curves can be replotted as a function of the strength of inactivation of MSTd (red) or VIP (blue) yielding behavioural predictions for partial inactivation of the areas. **(D)** Choice correlations (CC) of neurons in MSTd (blue) and VIP (red), before and after inactivation of VIP and MSTd respectively. Again the results following MSTd inactivation do not discriminate the two information models, but for VIP inactivation the predictions differ, showing increased CCs for the extensive information model and decreased CCs for the limited information model. Slopes of the lines correspond to $\zeta_M$ and $\zeta_V$ in **Eq (25)**, and shaded regions indicate ±1 s.d. of uncertainty.

https://doi.org/10.1371/journal.pcbi.1006371.g008

Likewise, the variance of the estimate also depends on the size of the neural recording. Although we extrapolated our data to larger populations by resampling from a set of about 100 neurons recorded from each area, our results are not attributable to the limited size of the recording (S14 Fig). We also extended our model to account for the fact that the two brain areas may have only been partially inactivated by Muscimol, and found that our conclusions hold under a wide range of partial inactivations (S7 Text; S15 Fig). Finally, we assumed that inactivation leaves responses in the un-inactivated area unaffected, as would be the case in a purely feedforward network model. While an exhaustive treatment of recurrent networks is beyond the scope of this work, we find that our conclusions can still hold at equilibrium if the above assumption is compromised by certain types of recurrent connections between MSTd and VIP (S8 Text; S16 Fig).

**Comparison of the two decoding strategies.** We inferred decoding weights in the presence of two fundamentally different types of noise, the extensive information model and the limited information model. Both of these decoders could account for the behavioural effects of selectively inactivating either MSTd or VIP, albeit with very different readout schemes. For the extensive information model, neurons in area VIP were weighted more heavily than optimal,

and vice-versa in the presence of information-limiting noise (Table 1, Fig 8A). Why do the two models have such different weightings? Both noise models have larger noise in VIP than MSTd, but differ in correlations between the two areas. In the extensive information model, the interareal correlations must be nearly zero to be consistent with behavioural data (Fig 5G), and the neuronal weights in VIP must be high to account for the high CCs. In the limited information model, the significant interareal correlations explain the large CCs in VIP, even with a readout mostly confined to MSTd.

How could such fundamentally different strategies lead to the same behavioural consequences? For a given noise model, an optimal decoder achieves the lowest possible behavioural threshold by scaling the weights of neurons in the two areas according to a particular optimal ratio $a_M/a_V$. Ratios that are either smaller or larger than this optimum will both result in an increase in the behavioural threshold due to suboptimality. This produces a *U-shaped* performance curve. Under certain precise conditions, complete inactivation of one of the areas will leave behavioural performance unchanged, exactly on the other side of the optimum. This is the case for VIP according to the extensive information model (**Fig 8B – top**). On the other hand, if the weight is already too small to influence behaviour then inactivation may not appreciably change performance, as demonstrated by the limited information model (**Fig 8B – bottom**).

**Model predictions.** According to the extensive information model, the brain loses almost all of its information by poorly weighting its available signals. Moreover, even beyond this poor overall decoding, the model brain gives VIP too much weight. As a consequence, this model makes a counterintuitive prediction that gradually inactivating VIP should *improve* behavioural performance! A hint of this might already be seen in Fig 2D and S5B Fig for the vestibular condition (both 0 and 12 h), although the difference was not statistically significant. Beyond a certain level of inactivation, as the weight decreases past the optimal scaling of the two areas, performance should worsen again (Fig 8C – top). According to the extensive information model, the brain just so happens to overweight VIP under normal conditions by about the same amount as it underweights VIP after inactivation. Suboptimal decoding in the limited information model has the opposite effect, giving too little weight to VIP, while overweighting MSTd. However, according to this model, the available information in VIP is small, because when MSTd is inactivated the behavioural thresholds are substantially worse (Fig 8C – bottom). Thus the suboptimality due to underweighting VIP is mild (around 80% in both visual and vestibular conditions, as described above), and the predicted improvement following partial MSTd inactivation is negligible as gradual inactivation quickly shoots past the optimum. Graded inactivation of brain areas can be accomplished by varying the concentration of muscimol, as well as the number of injections. In fact, we have previously reported that behavioural thresholds increase gradually depending on the extent of inactivation of area MSTd [22]. Unfortunately, those results do not distinguish the two models, as there is no qualitative difference between the model predictions for partial MSTd inactivation (Fig 8C, red). Future experiments involving graded inactivation of VIP should be able to distinguish between the models due to the stark difference in their behavioural predictions.

The decoding strategies implied by the two models also have different consequences for how CCs should change during inactivation experiments (**Methods, Eq (25)**). According to the extensive information model, VIP and MSTd are nearly independent, and both are decoded, so inactivating either area must scale up neuronal CCs in the other area (**Fig 8D – top**). In the limited information model, inactivating either area produces no significant changes in the other's CCs (**Fig 8D – bottom**). This effect has different origins for MSTd and VIP. Although inactivating MSTd confines the readout to VIP, it also eliminates the high-variance noise components that VIP shared with MSTd: these two effects approximately cancel

leaving CCs in VIP essentially unaffected. The results of VIP inactivation are simpler to understand: CCs in MSTd do not change much because VIP has little influence on behaviour to begin with.

## Discussion

Several recent experiments show that silencing brain areas with high decision-related activity does not necessarily affect decision-making[16–19]. To explain these puzzling results, we have developed a general, unified decoding framework to synthesize outcomes of experiments that measure decision-related activity in individual neurons and those that measure behavioural effects of inactivating entire brain areas. We know from the influential work of Haefner et al [14] how the behavioural impact (*readout weights*) of single neurons relates to their decision-related activity (*choice correlations*) in a standard feedforward network. We built on this theoretical foundation by adding three new elements that helped us relate the influence of multiple brain areas to both the magnitude of choice correlations, and the behavioural effects of inactivating those areas.

First, we have generalised their readout scheme to include multiple correlated brain areas by formulating the output of the decoder as a weighted sum of estimates derived from decoding responses of individual areas. In this scheme, the weight scales of individual estimates can be readily identified as the scaling of neuronal weights in the corresponding areas, providing a way to quantify the relative contribution of different brain areas. Second, we *postulated* that readout weights are mostly confined to a low-dimensional subspace of neural response that carries the highest response covariance, in both the extensive and limited information models. This postulate was instrumental to developing a theory of decoding that focused on the relationship between the overall scales of choice-related activity and neuronal weights, in lieu of their fine structures. Besides its mathematical simplicity, the resulting coarse-grained formulation confers an important practical advantage in that we can apply it without precisely knowing the fine structure of response covariance. Third, we used a straight-forward relation between behavioural threshold and the variance of the decoder to explicitly link the relative scaling of weights across areas to the behavioural effects of inactivating them.

Our theoretical result linking the behavioural influence of brain areas to their CCs and inactivation effects (**Eqs (20)** and **(21)**) is applicable only when neuronal weights within each area are mostly confined to the leading dimension of their response covariance. Although this requirement looks stringent, it is needed to explain the high CCs seen in experiments[15]. This claim might appear to be at odds with the fact that some earlier studies successfully predicted CCs that plateaued close to experimental levels using pooling models that did not explicitly take care of the above confinement[6,9]. However, a closer examination revealed that these studies used a scheme in which each decision was based on the average response of neuronal pools that were all uniformly correlated, a combination of model assumptions that in fact satisfies our requirement. Similar explanations apply to other simulation studies that used support-vector machines or alternative schemes that inadvertently restricted decoding weights to low-frequency modes of population response where shared variability was highest[12,30]. Thus our postulate is fully compatible with earlier work and in fact points to a more general class of models that can be used to describe the magnitude of CCs in those data.

Recent experiments show that reversibly inactivating area VIP in macaque monkeys does not impair animals' heading perception, despite the fact that responses of VIP neurons are strongly predictive of perceptual decisions[18,21]. In contrast, inactivating MSTd does adversely affect behaviour even though MSTd neurons exhibit much weaker correlations with choice[22,23]. Assuming that both areas contribute to decisions, we used our framework to

infer decoding strategies that could account for these experimental results. Surprisingly, the data were consistent with two different schemes – *overweighting* or *underweighting* of VIP – depending on whether information was *extensive* or *limited*. A major implication of the finding from the extensive information model is that if a causal test of function (e.g., inactivation) reveals no impairments, it does not disprove that a brain area contributes to a task. The limited information model on the other hand suggests that area VIP is indeed of very little use to heading perception. In spite of this difference, both models share a basic attribute, namely, that decoding is suboptimal (although to very different extents, as discussed in the next section). Therefore, our analysis reveals that the observed discrepancy between decision-related activity and effects of inactivation is not peculiar, and is actually expected from systems that integrate information across brain areas in a suboptimal fashion. The nature of this suboptimality can be understood intuitively by drawing an analogy to cue combination. Imagine there are two cues $x$ and $y$, and you use a suboptimal strategy in which a larger weight is allocated to the less reliable cue $y$. If $y$ is removed thereby forcing you to rely completely on $x$, then your behavioural precision might not change very much if the reduction in information from losing $y$ is offset by the gain in information from $x$. On the other hand, if you mostly ignored $y$ to begin with, then once again you will be unaffected by its removal. Either "too much" or "too little" weighting of a brain area can lead to suboptimal performance, both in a way that leaves the behavioural threshold largely unaltered following complete inactivation of that area.

## Decoding is suboptimal, but just how bad?

Although both models were suboptimal to some degree, the overwhelming distinction between them is the efficiency they imply for neural computation, where efficiency is the ratio of decoded information to available information. The efficiency of the limited information model is around 80%, independent of population size $N$. In contrast, the extensive information model encodes information that grows with $N$, while decoding is restricted to the least informative dimensions of neural responses. These decoders extract only a tiny fraction of the available information, resulting in an efficiency that falls inversely with $N$. For a modest-sized population of 1000 neurons, the efficiency is already less than 1%. Thus, the conventional model of correlated noise (with extensive information) is radically suboptimal, whereas the limited information model extracts an impressive fraction of what is possible, limited largely by noise.

It has previously been argued that the key factor that limits behavioural performance in complex tasks is suboptimal processing, not noise[39]. However, in simple tasks involving binary choices, and in areas in which most of the available information can be linearly decoded, it is unclear why the behaviour of highly trained animals should be so severely undermined by suboptimality. Moreover, radical suboptimality of the kind described here for the extensive information model implies tremendous potential for learning, as the neural circuits can continually optimize the computation by tuning the readout to more informative dimensions. This is hard to reconcile with the observation that behavioural thresholds in a variety of perceptual tasks typically saturate within a few weeks of training in both humans and monkeys [29,40–42]. In the presence of information-limiting noise, however, learning can only do so much, and performance must saturate at or below the ideal performance. Therefore, we regard the limited information model as a much more likely explanation of our data, for otherwise one would need to posit that cortical computations discard the vast majority of available information. Note that suboptimal cortical computation might still account for information loss in the limited information model, as opposed to neural noise[39], but this information loss is now much more modest, probably around 20%.

A direct way to tell the two models apart would be to measure the structure of noise correlations. Unfortunately, this is not straightforward, because the differences between noise models giving extensive or limited information can be quite subtle[20]. In fact, there can be a whole spectrum of subtly different noise models with different information contents, lying between the two models that we have considered here. Therefore, a more accurate technique to determine the information content (which, after all, is a major reason why we care about noise correlations) is simply to record from hundreds of neurons simultaneously, and then decode the stimulus. This will provide a lower bound on the information available in the neural population. One can then compare the resultant population thresholds with the behavioural threshold to determine how suboptimal the decoding needs to be to account for behaviour. Eventually, we expect this strategy will be successful, but it will require advances in recording technology to be viable in the target brain areas. Meanwhile, by examining the key properties of the decoding strategy implied by the two models, we identified distinct predictions that are testable without large-scale simultaneous recordings. Specifically, they involve fairly simple experiments such as graded inactivation of VIP, and measurement of CCs in either VIP or MSTd while the other area is inactivated (**Fig 8**). Future experiments will test each of these predictions to provide novel evidence about the information content and decoding strategy used by the brain.

## Limitations of the framework and possible extensions

Similar efforts to deal with outcomes of correlational and causal studies using a coherent framework are rarely undertaken, despite their significance. To our knowledge, there is only one instance where this has been attempted before[43]. In that work, the authors used a recurrent network model with mutual inhibition between populations[44,45] to reconcile choice-related activity and the effect of silencing neurons. Although their study was similar to ours in spirit, their goal was different. They showed that inactivation just before a decision, when activity was highly correlated with the choice, had less impact on the behaviour than inactivation near the stimulus onset. This addresses a *temporal*, as opposed to a *spatial*, dissociation between correlation and causation, so a model with recurrent connectivity was essential to explain their findings. In contrast, we wanted to account for the discrepancies between measures of correlation and causation across brain areas. This latter phenomenon is entirely within the realm of standard feedforward network models in which both populations causally contribute, rather than compete to drive behaviour, and differ only in terms of the relative strength of their contributions.

Time-varying weights have been shown to better predict animals' choice in certain tasks [46], and psychophysical kernels are sometimes skewed towards one end of the trial[47,48], suggesting that decoding could also be suboptimal in time. Consistent with suboptimal integration, choice correlations in our task peak before the end of the trial, even though new evidence is still available (**Fig 7B**). Such temporal weighting of information would naturally arise from recurrent connectivity, which is beyond the scope of this work. But it can also originate in feedforward networks, possibly through a gating mechanism that blocks the integration of neural responses beyond a certain time.[32]

Other studies have considered that choice-related activity might arise from decision feedback[47,49,50]. Indeed, pure decision feedback to an area would create apparent sensitivity to sensory signals, even in the absence of direct feedforward input to the target neurons [47,49,50]. In such a case, neural sensitivity to the stimulus would then be precisely equal to the animal's sensitivity. In the absence of other sources of variability, response fluctuations would be perfectly correlated with fluctuations in the fed-back choice, producing choice correlations of 1. Of course there would be additional variability in the neural responses, and this

would dilute both the choice correlations and neural tuning by equal amounts, giving rise to measured CCs that should match the optimal CCs (**Eq (4)**). Even if there are other feedforward sensory components to the neural responses, direct decision feedback will pull the choice correlations toward this optimal prediction. Thus, simple decision feedback cannot account for the pattern of CCs observed in our VIP data, which are two to three times larger than predicted from optimal inference or direct decision feedback (**Fig 3**). Conversely, as we demonstrated through supplementary modeling, adding feedback or recurrent connections may not affect the suboptimal readout weights inferred using our scheme, even when those connections modulate responses along the decoded dimensions (**S16 Fig**). Nevertheless, future expansions of our work should account for more general recurrent connectivity to study how neural circuits simultaneously integrate information across space and time. In particular, recurrent networks also include decision feedback as a special case, and might help test alternative theories on the origins of choice correlations[1,47].

Finally, while VIP inactivation did not impair heading discrimination, MSTd inactivation partially impaired the animal's ability to perform the task. The fact that MSTd inactivation did not completely abolish performance cannot be accounted for by our two-population models unless the inactivation was only partial and/or VIP is read out to some degree. Additionally, we cannot exclude the possibility that VIP is merely correlated with behaviour and that a third brain area besides MSTd contributes some task-relevant information. In fact, both of our models actually predict a somewhat bigger deficit following MSTd inactivation (**Figs 5C and 6A**) than is observed experimentally (**Fig 1B**). This highlights the importance of ultimately extending coding models to include more than two brain areas.

As neuroscience moves towards 'big data', there is a greater need for theoretical frameworks that can help discern simple rules from complex multi-neuronal activity[51]. We believe our work responds to this challenge and, despite its limitations, takes us closer to bridging the brain-behaviour gap for binary-decision tasks.

## Methods

### Ethics statement

All surgical and experimental procedures were approved by the Institutional Animal Care and Use Committees at Washington University and Baylor College of Medicine, and were performed in accordance with institutional and National Institutes of Health (NIH) guidelines.

### Relation between behavioural threshold and weight scaling factors

Behavioural threshold $\vartheta$ is proportional to the square root of the decoder variance (with proportionality of 1 for threshold of 68% correct), so $\vartheta^2 = \mathbf{w}^{\mathrm{T}}\Sigma\mathbf{w}$. If decoding is confined to the subspace of leading eigenmodes $\mathbf{u}^x$ of $\Sigma$ spanned by neurons within each population $x$, then $\mathbf{w}_x = \mathbf{u}^x/(\boldsymbol{f}_x'^{\mathrm{T}}\mathbf{u}^x)$ where the constant of proportionality ensures unbiased decoding from that population. In this case, the behavioural threshold can be expressed purely in terms of weight scaling factors and the variance originating from noise within the noise modes as (**S3 Text**):

$$\vartheta^2 = \boldsymbol{a}^{\mathrm{T}}E\boldsymbol{a} = a_x^2\varepsilon_{xx} + a_y^2\varepsilon_{yy} + 2a_xa_y\varepsilon_{xy} \tag{22}$$

where $E = \varepsilon_{xy}$ is the covariance matrix of the noise decoded from populations $x$ and $y$. Thresholds following inactivation can be determined by setting the weight scaling factor for the inactivated areas to zero. In the case of two populations, this yields $\vartheta^2_{-x} = \varepsilon_{yy}$ and $\vartheta^2_{-y} = \varepsilon_{xx}$.

## Subjects and behavioural task

Six adult rhesus monkeys (A, B, C, J, S, U, and X) took part in various aspects of the experiments. Three animals were employed in each of the MSTd (C, J and S) and VIP (X, B and J) inactivation experiments. Two animals provided the neural data from each brain area (A and C for MSTd; C and U for VIP). All animals were trained to perform a heading discrimination task around psychophysical threshold. In each trial, the subject experienced a real or simulated forward motion with a small leftward or rightward component (angle $s$, **Fig 1A**). Subjects were required to maintain fixation within a 2x2˚ electronic window around a head-fixed visual target located at the center of the display screen. At the end of each 2-s trial, the fixation spot disappeared, two choice targets appeared and the subject made a saccade to one of the targets to report his perceived heading relative to straight ahead. Nine logarithmically spaced heading angles were tested (0˚, ±0.5˚, ±1.3˚, ±3.5˚, and ±9˚ for monkeys A and J, 0˚, ±1˚, ±2.5˚, ±6.4˚, and ±16˚ for monkeys B, C, S and U), including the ambiguous case of straight ahead motion ($s = 0$˚). These values were chosen to obtain near-maximal psychophysical performance while allowing neuronal sensitivity to be estimated reliably for most neurons[21,23]. Subjects received a juice reward for indicating the correct choice. For trials in which the ambiguous heading was presented, rewards were delivered randomly on half of the trials. The experiment consisted of three randomly-interleaved stimulus conditions (vestibular, visual, and combined). In the vestibular condition, the monkey was translated by a motion platform while fixating a head-fixed target on a blank screen. In the visual condition, the motion platform remained stationary while optic flow simulated the same range of headings. Under the combined condition, both inertial motion and optic flow were provided. Each of the 27 unique stimulus conditions (9 heading directions × 3 cue conditions) was repeated at least 20 times, for a total of 540 discrimination trials per recording session. Identical stimuli and trial structure were employed during both neural recordings and inactivation experiments.

## Neural recordings

Activity of single neurons in areas MSTd and VIP was recorded extracellularly using epoxy-coated tungsten microelectrodes (impedance of 1–2 MΩ). Area MSTd was located using a combination of magnetic resonance imaging (MRI) scans, stereotaxic coordinates (~15 mm lateral and ~3–6 mm posterior to AP-0), white/gray matter transitions, and physiological response properties. In some penetrations, electrodes were further advanced into the retinotopically organized area MT[23]. Most recordings concentrated on the posterior/medial portions of MSTd, corresponding to more eccentric, lower hemifield receptive fields in the underlying area MT. To localize area VIP, we first identified the medial tip of the intraparietal sulcus and then moved laterally until there was no longer directionally selective visual response in the multiunit activity, as described in detail previously[21].

## Estimation of behavioural and neuronal thresholds

Behavioural performance was quantified by plotting the proportion of 'rightward' choices as a function of heading (the azimuth angle of translation relative to straight ahead). Psychometric data were fit with a cumulative Gaussian function with mean $\mu$ and standard deviation $\vartheta$, and this standard deviation defined the psychophysical threshold, corresponding to 68% correct performance ($d' = 1$, assuming no bias, i.e. $\mu = 0$).

For the analysis of neuronal responses, we used the linear Fisher information $J$ which is simply a measure of the signal-to-noise ratio: signal power divided by noise power. The linear Fisher Information captures all of the Fisher information in responses generated from the exponential family with linear sufficient statistics. Its inverse is exactly equal to the variance of

an unbiased, locally optimal linear estimator (for differentiable tuning curves and nonsingular noise covariance). We defined the square root of this variance (i.e. the standard deviation of the estimator) to be the neuronal discrimination threshold, which corresponds to 68% accuracy in binary discrimination. This threshold can be obtained directly from the neuron's tuning curve and noise variance as follows:

$$\vartheta_k = \frac{1}{\sqrt{J_k}} = \frac{\sigma_k}{f_k'} \tag{23}$$

where $\vartheta_k$ and $J_k$ are the threshold and linear Fisher information[52] for neuron $k$, $f_k'$ is the derivative of the neuron's tuning curve at the reference stimulus (0°), and $\sigma_k^2$ is the variance of the neuronal response for that stimulus. Neuronal thresholds computed using the above definition were very similar to those computed using a traditional approach based on neurometric functions constructed from the responses of the recorded neuron and a presumed 'antineuron' with opposite tuning[53] (**S4 Fig**).

## Estimation of choice correlation

To quantify the relationship between neural responses and the monkey's perceptual decisions, we first computed choice probabilities (CP) using ROC analysis[54]. For each heading, neural responses were sorted into two groups based on the choice that the animal made at the end of each trial. In previous studies, the two choice groups were typically related to the preferred and non-preferred stimuli for a given neuron[21,23]. In this study, in order to appropriately compare different neurons in a population code, the two choice groups were simply rightward and leftward choices; hence, CPs may be greater than or less than 1/2. ROC values were calculated from these response distributions, yielding a CP for each heading, as long as the monkey made at least 3 choices in favor of each direction. To combine across different headings, we computed a grand CP for each neuron by balanced $z$-scoring of responses in different conditions, which combines $z$-scored response distributions in an unbiased manner across conditions, and then performed ROC analysis on that combined distribution[55]. The CPs were then converted to choice correlations according to $C_k \approx \frac{\pi}{\sqrt{2}}\left(CP_k - \frac{1}{2}\right)$ (refs. [14,15]) where $CP_k$ and $C_k$ are the choice probability and choice correlation of neuron $k$ respectively (**S1 Text**). Due to the convention we chose for computing CPs, the resulting choice correlation could be positive or negative depending whether a neuron predicted *rightward* choices by increasing or decreasing its response relative to reference stimulus. For an optimal decoder, the sign of a neuron's choice correlation should match the sign of the derivative of its tuning curve, so we modified the definition of ref.[15] (**Eq (4)**) to accommodate our sign convention, yielding $C_{k,\mathrm{opt}} = \mathrm{sgn}(f_k')\vartheta/\vartheta_k$ where sgn denotes the signum function.

There were neurons in both MSTd and VIP whose choice-related activity during the visual condition is anticorrelated with their signal-related activity[21,23]. Further analysis showed that heading preferences of these neurons during visual and vestibular conditions differed. Therefore the analysis of data collected during the visual condition presented in the supporting material included only the subset of recorded neurons that had similar heading preferences as in the vestibular condition[23] (MSTd: 66/129 neurons; VIP: 63/88 neurons).

## Noise covariance of extensive information model

Pairwise neuronal recordings carried out separately in areas VIP and MSTd were used to estimate noise correlations between pairs of neurons, $R_{ij} = \mathrm{Corr}(r_i, r_j | s = 0)$, where $r_i$ and $r_j$ are the responses of neurons $i$ and $j$, and correlation coefficients were computed by averaging over

trials with headings near 0˚. The same recordings were used to compute signal correlations, $R_{ij}^{sig} = \mathrm{Corr}(f_i, f_j)$, where $f_i$ and $f_j$ are the tuning curves of neurons $i$ and $j$, and the correlation coefficients were computed by averaging over a uniform distribution of headings in the horizontal plane. The typical noise correlations, $\bar{R}$, were then modeled as linearly proportional to the signal correlations (**Eq (8)**). The slope of the relation was much steeper in VIP than MSTd [21]. For the vestibular condition, slopes were found to be $m_M = 0.19 \pm 0.08$ and $m_V = 0.70 \pm 0.16$ within MSTd and VIP respectively, and for the visual condition they were $m_M = 0.12 \pm 0.09$ and $m_V = 0.50 \pm 0.14$. The above fits determined the average relationship between noise and signal correlations, but there was considerable diversity around this trend. To emulate this diversity, we used a technique similar to the one proposed in ref. [31]. Specifically, we sampled correlation coefficient matrices $R$ from a Wishart distribution with a mean matrix $\bar{R}$ given by **Eq (8)** and the fitted slope $m$, and rescaled them to ensure $R_{ii} = 1$. The number of degrees of freedom for the Wishart distribution was adjusted so sampled matrices had the same uncertainty in slope $m$ as the data when subjected to the same fitting procedure. Covariance matrices were generated by scaling the correlation coefficients by the standard deviations for each neuron. Model variances were set equal to the mean responses, so the standard deviation of neuron $i$ is $f_i^{1/2}$. Thus the covariance $\Sigma$ is related to correlation coefficients $R$ by $\Sigma_{ij} = R_{ij}\sqrt{f_i f_j}$.

Correlations between responses of MSTd and VIP neurons were not measured experimentally, so the slope $m_{MV}$ of any linear trend relating noise and signal correlations between the two areas was not known. We explored different possibilities by varying $m_{MV}$ according to:

$$m_{MV} = k\sqrt{m_M m_V} \tag{24}$$

where $k \in [0,1)$. Each value of $k$ produced correlation between areas with magnitude $\varepsilon_{MV}$ which was expressed as $\varepsilon_{MV} = \gamma\varepsilon_{MM}$.

## Noise covariance of limited information model

If the information reaching MSTd ($M$) and VIP ($V$) is not perfectly redundant across the populations, then the resulting covariance matrix will be of the form given by **Eq (13)** where $M$ and $V$ take the places of $x$ and $y$. The resultant covariances $\varepsilon_{MM}$, $\varepsilon_{VV}$, and $\varepsilon_{MV}$ are difficult to determine even with large-scale recordings since their magnitudes may be very small compared to the magnitude of noise in $\Sigma$. Nevertheless, we know that for large populations, the behavioural threshold will be dominated by the magnitude of information-limiting correlations. Specifically, they are related through the relative scaling of decoding weights in **Eq (22)**. Consequently, we can determine $\varepsilon_{MM}$ and $\varepsilon_{VV}$ from behavioural thresholds following inactivation using $\varepsilon_{MM} = \vartheta_{-V}^2$ and $\varepsilon_{VV} = \vartheta_{-M}^2$. We can then use **Eq (22)** in conjunction with **Eq (21)** to determine both the ratio $a_M/a_V$ of scaling factors and the magnitude of correlation between populations $\varepsilon_{MV} = \gamma\varepsilon_{MM}$.

## Effects of inactivation on choice correlations

Complete inactivation of one of the areas will affect neuronal choice correlations in the non-inactivated area. If $\mathbf{C}_x$ and $\tilde{\mathbf{C}}_x$ denote the choice correlations of neurons in area $x$ before and after inactivation of $y$, then it can be shown that $\tilde{\mathbf{C}}_x = \zeta_x \mathbf{C}_x$ and similarly $\tilde{\mathbf{C}}_y = \zeta_y \mathbf{C}_y$ where scalars $\zeta_y$ and $\zeta_y$ are (S9 Text):

$$\zeta_x = \frac{1}{\beta_x}\frac{\vartheta_{-y}}{\vartheta};\ \zeta_y = \frac{1}{\beta_y}\frac{\vartheta_{-x}}{\vartheta} \tag{25}$$

where $\beta_x$ and $\beta_y$ are the multipliers that relate the observed and optimal patterns of neuronal choice correlations in areas $x$ and $y$. The above equation implies that choice correlations in the active area will increase by a factor proportional to the behavioural effect of inactivating the other area. Intuitively, this is because inactivating an area that was very important for behaviour will dramatically increase the burden on the active area, leading to an increase in the magnitude of choice-related activity.

## Supporting information

**S1 Fig. Choice correlations decrease with the number of decoded modes. (A)** Tuning functions $f_i(s)$ (left) and covariance matrix $\Sigma$ (right) of a subset of model neurons used in this simulation. The stimulus $s \in (-\pi,+\pi]$ was a circular variable and tuning followed a von Mises function: $f_i(s) = b_i + h_i e^{\kappa_i \cos(s-s_i)}$ where baseline and height $b_i$ and $h_i$ were drawn from Poisson distributions $b_i \sim Poiss(\bar{b})$ and $h_i \sim Poiss(\bar{h})$ with means $\bar{b}$ = 5 spikes/sec and $\bar{h}$ = 15 spikes/sec, tuning peakiness $\kappa_i$ was sampled from the rectified normal distribution $\kappa_i \sim |\mathcal{N}(1, 0.25)|$, and preferred stimulus $s_i$ was drawn from a uniform distribution. Covariance $\Sigma_{ij}$ between neurons $i$ and $j$ was $\Sigma_{ij} = R_{ij}\sqrt{f_i f_j}$ where noise correlation coefficient $R_{ij}$ was proportional to signal correlation (**Eq (8)**) with a proportionality of 0.2. **(B)** Neurons were linearly decoded by confining readout weights to the leading $p$ eigenmodes of the covariance. Weights were always chosen to be optimal within the decoded subspace, and $p$ was varied from 1 to $N$ where $N$ = 512 denotes the population size. The root-mean-squared choice correlation $C_{\mathrm{RMS}}$ over all neurons decreases with $p$: for this model population, it drops by an order of magnitude already for $p$ = 2. Inset shows $C_k$ of each neuron for two example cases. **(C)** Choice correlations tend to decrease with population size when all modes are decoded optimally (gray: $p = N$), but remain insensitive to population size when only the leading mode is decoded (black: $p$ = 1).
(PDF)

**S2 Fig. Recovering the true values of the decoder scaling factors in simulated neural populations.** In this demonstration, 6 populations with information-limiting noise are each manipulated by a random multiplicative inactivation factor. We successfully recover decoder scalings (left) and population noise covariance (right) using behavioural thresholds and choice correlation slopes during these inactivation experiments by numerically solving **Eqs (18)** and **(19)**.
(PDF)

**S3 Fig. Inactivation effects may not reflect relative influence of brain areas on behaviour.** Consider two populations $x$ and $y$ with relative scaling of neuronal weights $a_x$ and $a_y$. These scalings depend not only on the post-inactivation thresholds ($\vartheta_{-x}$ and $\vartheta_{-y}$) but also on the magnitude of their choice correlations ($\beta_x$ and $\beta_y$) according to **Eqs (20)** and **(21)**. The two panels illustrate the relative choice correlation magnitudes ($\beta_x/\beta_y$, color) for uncorrelated populations (**Eq (20)**) and correlated populations (**Eq (21)**), as a function of the scaling ratio $a_x/a_y$ and the inactivation ratio ($\vartheta^2_{-x}/\vartheta^2_{-y}$). For simplicity, here we assume that $\beta_y$ = 1, so $\beta_x/\beta_y$ = 1 corresponds to optimal decoding. (**A**) For systems in which the two populations are uncorrelated ($\varepsilon_{xy}$ = 0), the scaling ratio $a_x/a_y$ is directly proportional to inactivation ratio $\vartheta^2_{-x}/\vartheta^2_{-y}$. Nonetheless the slope of this relationship depends on the ratio of choice correlation magnitudes $\beta_x/\beta_y$ (isochromatic contours), so a population with a larger weight could produce a smaller deficit upon inactivation, or vice-versa (black asterisks). Inactivation effects exactly match the ratio of scalings (e.g. black open circle on the main diagonal) only if decoding is optimal (black dashed line). (**B**) When the populations are correlated, the scaling ratio is no longer proportional to the inactivation ratio. Instead, their relationship is nonlinear (black dashed line), and the two

ratios may not match even if decoding happens to be optimal (e.g black open circle). In other words, the change in behavioural threshold does not match how much each area is decoded. Here cross-population correlation $\varepsilon_{xy}$ is $\sqrt{\varepsilon_{xx}\varepsilon_{yy}}/2$ for illustration.
(PDF)

**S4 Fig. Direct and conventional methods yield similar neuronal thresholds.** Each neuron's threshold was estimated in two ways—directly as the inverse square-root of its Fisher information at $s = 0$ (Methods – **Eq (23)**), or using a traditional approach by constructing a neurometric function. The latter approach used ROC analysis to compute the ability of an ideal observer to discriminate between two oppositely-directed headings (e.g., –6.4˚ vs. +6.4˚) based solely on the firing rate of the recorded neuron and a presumed 'antineuron' with opposite tuning[1]. ROC values were plotted as a function of heading, resulting in neurometric functions that were fit with a cumulative Gaussian function. Neuronal threshold was then defined as the standard deviation of the fitted Gaussian, but increased by a factor of $\sqrt{2}$ to adjust for the extra information from the antineuron. This $\sqrt{2}$ adjustment arises because a decision based on a neuron-antineuron pair has twice the signal amplitude but also twice the noise variance, compared to a single neuron and a fixed, noiseless 0˚ reference. Note that this factor of $\sqrt{2}$ differs from past studies[2] that assumed a noisy 0˚ reference heading and thus corrected by a factor of 2. (**A**) The two methods yielded very similar estimates for vestibular thresholds across neurons in both MSTd (blue, Pearson's correlation $r = 0.55$, $p = 4 \times 10^{-11}$) and VIP (red, $r = 0.31$, $p = 5 \times 10^{-3}$). (**B**) Similar results were found for visual thresholds: MSTd (blue, $r = 0.65$, $p = 3 \times 10^{-9}$) and VIP (red, $r = 0.87$, $p = 1 \times 10^{-20}$). For these comparisons, we omitted a small subset of insensitive neurons (Vestibular: 4/129 MSTd neurons and 7/88 VIP neurons, Visual: 1/129 MSTd neurons and 5/88 VIP neurons) with extremely large thresholds (>300˚).
(PDF)

**S5 Fig. (A) Choice correlations of VIP neurons.** Neural recordings were carried out in a separate monkey X prior to inactivation of area VIP, while he performed a heading discrimination task whose structure was identical to that described in Methods in all regards, except each trial lasted only 1s instead of 2s. Similar to those in monkeys C and U, neuronal choice correlations in area VIP are proportional to but greater than those expected from optimal decoding of these neurons during both vestibular (top) and visual (bottom) heading discrimination tasks. The 95% CI of slopes $\beta_V$ were found to be [1.9 2.9] and [1.2 1.8] for the vestibular and visual conditions respectively. **(B) Behavioural effects of VIP inactivation.** *Left*: Discrimination thresholds at different times (different shades of blue) following inactivation of VIP, for all seven experiments conducted on monkey X. Thresholds obtained in a single experimental session are connected by a line. Across experiments, inactivating area VIP failed to elicit significant changes in either the vestibular or visual conditions. The behaviour of this monkey was tested 36 hours following inactivation in only 3 of the 7 experiments. *Right*: Psychometric functions at different times during inactivation of area VIP, averaged across experiments, for the vestibular (top) and visual (bottom) conditions. Behavioural thresholds computed from the psychometric functions at different times are shown in the bottom panels. None of the comparisons were significant (Wilcoxon rank-sum test, significance-level of $p = 0.05$). Error bars indicate standard error of the mean.
(PDF)

**S6 Fig. Pattern of choice correlations in individual animals.** Experimentally measured choice correlations ($C_k$) of neurons in MSTd (blue) for both the vestibular (top) and the visual (bottom) condition are close to optimal predictions ($C_{k,\mathrm{opt}}$), those of VIP neurons are systematically greater (red). This observation holds individually in each monkey. Solid black lines

correspond to the best linear fit. Vestibular data in monkeys C and U are replotted from Ref. [15] with different sign convention (see **Methods**). Monkey A was used only for MSTd recordings, and monkeys U and X only for VIP.
(PDF)

**S7 Fig. Noise and signal correlations.** Pairs of neurons within MSTd (blue; $n$=127 pairs) and VIP (red; $n$=139 pairs) were recorded when the animal experienced self-motion in various directions based on either vestibular (left) or visual (right) cues. For each pair of neurons $i$ and $j$, correlated variability in the firing rates across trials (*noise correlation* $R_{ij}^{\text{noise}}$) is plotted against correlated variability in the average firing rates across stimuli (*signal correlation* $R_{ij}^{\text{sig}}$). The relationship between signal and noise correlation was fit to a linear model (**Eq (8)**) separately for each area, represented here using straight lines. Shaded areas correspond to 95% confidence intervals of the resulting fits.
(PDF)

**S8 Fig. Noise along the leading modes of covariance substantially influences choice.** Any readout weight $\mathbf{w}$ can be expressed as a linear combination of the eigenvectors $\mathbf{u}_p$ of the response covariance as $\mathbf{w} \propto \sum_p a_p \mathbf{u}_p$ where the constant of proportionality is chosen to ensure unbiased decoding. Coefficients magnitudes $|a_p|$ indicate how much the different eigenmodes $p$ contribute to behavioural choice. To assess the specific contribution of the leading mode from MSTd and VIP, we considered three different cases: optimal decoding of response along all available modes, a decoder confined to the leading eigenmode in each area, and a spectrum of decoders in between the two extremes. We decoded MSTd & VIP responses separately in all cases using covariance $\boldsymbol{\Sigma}$ specified by the extensive information model (**Fig 4A** – *left*), and examined the average magnitude of choice correlations across all neurons in each case. **(A) Optimal decoding of all modes.** The pattern of coefficients $a_p$ of the optimal decoder of MSTd (blue) and VIP (red) responses. For clarity, only the coefficients corresponding to the leading 40 modes are shown. Evidently, the leading mode has little influence on the decoder output as seen from the magnitude of coefficient $a_1$. **(B)** The average choice correlation, quantified as the root-mean squared (RMS) choice correlations of the set of all neurons, decreases to ~0.01 even for the modest population size of $N$ = 1000 neurons. **(C,D) Decoding leading mode only.** Plotted as for A,B, but restricting the readout to one leading eigenmode. We forced the coefficients $a_p$ to zero for all $p \neq 1$, yielding $\mathbf{w} \propto \mathbf{u}_1$. Choice correlations implied by this decoder asymptote to about 0.2 and 0.4 for MSTd and VIP, values that are of the same order of magnitude as seen in the experiments. **(E) Varying weight on leading mode.** We tested whether the leading mode must contribute substantially to choice, in order to generate high choice correlations. To test this, we first parametrised the contribution of the leading mode as the fraction $\Phi$ of weight power it contributes to decoding, according to $\Phi = a_1^2 / \sum_{p=1}^{N} a_p^2$. To control $\Phi$, we simply manipulated the coefficients $a_p$ of the optimal decoder, first by setting the leading coefficient $\tilde{a}_1$ to $\sqrt{\Phi}$ and then rescaling all the remaining coefficients together so $\tilde{a}_p = a_p \sqrt{(1-\Phi) / \sum_{p=2}^{N} a_p^2}$. Weights obtained by this procedure resemble the optimal weight pattern except for the differences arising from the leading mode. We then systematically varied $\Phi$ from $1/N$ to 1 where the number of neurons $N$ was fixed to 1024 in this simulation. Choice correlations increase slowly with $\Phi$, and reach half-max at about $\Phi$ = 0.25 (dashed vertical line). **(F)** Influence of the leading mode on noise in the output increases much more rapidly with $\Phi$ than choice correlations do. For each value of $\Phi$, we computed the fraction $\xi$ of total noise variance that comes from the leading mode as $\xi = \tilde{a}_1^2 \lambda_1 / \sum_{p=1}^{N} \tilde{a}_p^2 \lambda_p$ where $\lambda_p$ denotes the eigenvalue of the $p^{\text{th}}$ mode. At $\Phi$ = 0.25, more than 95%

of noise propagated to the output is inherited exclusively from this mode (dashed vertical line).
(PDF)

**S9 Fig. Decoder inferred using the extensive information model – visual condition (compare to Fig 5B and 5C). (A)** Experimentally measured choice correlations ($C_k$) of individual neurons in MSTd (blue) and VIP (red) are plotted against the $i^{\text{th}}$ component $C^i_{k,\text{opt}}$ of choice correlations generated from optimally decoding the responses within the subspace of two leading principal components of noise covariance. When two populations are not correlated with each other, the two leading components of the global noise covariance correspond to the largest noise modes in each population separately. Consequently $C^1_{k,\text{opt}}$ and $C^2_{k,\text{opt}}$ correspond to optimal choice correlations in VIP and MSTd, respectively. **(B)** Performance (threshold) of a decoder with weights inferred from the subspace of two leading principal components of the noise covariance. The black and green lines indicate the performance of the inferred and optimal decoders within this subspace. Inactivating VIP is correctly predicted to have no effect on behavioural performance (blue), while MSTd inactivation increases the threshold (red). Shaded region indicates ±1 SEM.
(PDF)

**S10 Fig. Effect of the decoded subspace dimensionality on performance of the decoder inferred from choice correlations using the extensive information model.** Since decoding performance was nearly saturated at 256 neurons (Fig 5C), we fixed the size of the neural population at $N = 256$, and examined the behavioural threshold when varying the dimensionality of the decoded subspace. Decoding weights were inferred in the subspace spanned by a total of $p$ eigenvectors of the covariance matrix, using $p/2$ eigenvectors in both MSTd and VIP. The decoder continued to correctly predict the qualitative effects of inactivating MSTd and VIP beyond the 2-dimensional subspace considered in Fig 5, roughly until about $p = 22$ (vertical dashed line). Note that the threshold predicted by the optimal decoder within the restricted subspace (green) improves as more (informative) dimensions are included, while that of the inferred decoder worsens. Therefore, readout weights extract more noise than signal from these additional dimensions. This makes sense because if it the weights were instead tuned to decrease the variance in the estimate as more dimensions are added, they would no longer explain the large measured choice correlations. One reason why the experimental predictions of this model break down for large $p$ is that the predictions are only reliable in the regime of small $p$ where the effect of measurement noise is low. This is because the reliability of inferred decoding weights (and consequently also its predictions) is inversely related to the eigenvalue of the decoded mode, so reliability of the predictions worsens as $p$ increases (S6 Text).
(PDF)

**S11 Fig. Effect of interareal correlations on decoder inferred from choice correlations using the extensive information model.** *Left*: A representative covariance matrix when neurons in MSTd and VIP are mildly correlated through the leading noise modes ($\varepsilon_{xy} \approx 0.2\sqrt{\varepsilon_{xx}\varepsilon_{yy}}$). *Right*: In contrast to the observed effects of inactivation, the decoder inferred using the covariance on the left incorrectly predicted that inactivating VIP should reduce the behavioural threshold. This was unlike the decoder shown in Fig 5C that correctly predicted the effects of VIP inactivation when correlations between the two areas were zero on average.
(PDF)

**S12 Fig. Decoder inferred using the limited information model: visual condition.** **(A)** Like decoding in the presence of extensive information, this decoder is suboptimal (black vs green), and can account for the behavioural effects of inactivation. **(B)** Unlike decoding in the extensive information model, the efficiency of this decoder is quite high and insensitive to population size. Shaded areas represent ±1 SEM.
(PDF)

**S13 Fig. Readout weights for the visual condition do not vary drastically across time.** Neuronal thresholds **(A)** and choice correlations **(B)** were computed for each neuron across the duration of the trial using a 250ms moving window and averaged across neurons. Note that these readouts predict the choice based only on a single time window per data point, and do not perform a weighted sum of responses in multiple windows. Neuronal thresholds in both brain areas were comparable at all times, yet the choice correlations (CCs) differed between brain areas VIP and MSTd in a consistent manner over time. Although CCs in both areas peaked around the middle of the trial, those in VIP were proportionally larger at almost all times. **(C)** Consequently the slopes, $\beta = C_k/C_{k,\mathrm{opt}}$, that related observed and optimal choice correlations were generally greater in area VIP than in MSTd. **(D)** The readout weights inferred using the two models remain largely constant throughout the trial, and are qualitatively consistent with the conclusions drawn from our analyses presented in the main text: the extensive information model implies that area MSTd is underweighted, whereas the limited information model predicts the opposite. Symbols $a_M$ and $a_V$ denote scaling of readout weights of areas MSTd and VIP respectively. **(E)** Regression slopes are minimally affected by the length of the analysis window. Both observed neuronal choice correlations as well as those implied by optimal decoding of MSTd and VIP populations increased similarly with the length of the analysis window, leaving the regression slopes $\beta = C_k/C_{k,\mathrm{opt}}$ largely invariant with the window length for both VIP (red) and MSTd (blue). **(F)** The qualitative difference in the readout weights inferred using the two noise models are consistent across different lengths of analysis window. Error bars denote ±1 standard deviation.
(PDF)

**S14 Fig. Threshold saturation effects are not influenced by size of the dataset.** In the main text, we presented thresholds predicted by decoders inferred using the Extensive information (EI) (Fig 5C) and Limited information (LI) (Fig 6B) models. These thresholds were generated by extrapolating a limited dataset containing 129 and 88 neurons from MSTd and VIP respectively. However, those thresholds approached saturation only around 60-70 raising the possibility that those results might be sensitive to the exact number of neurons that were used for extrapolation. To test whether this was the case, we repeated all our analyses by considering only a fraction of the recorded neurons for extrapolation. **(A)** *Left*: Thresholds implied by the EI model obtained by extrapolating 50% of the neurons in our dataset ($n$=65/129 and 44/88 neurons in MSTd and VIP). Thresholds were found to asymptote to nearly the same value obtained by extrapolating the full dataset (compare with Fig 5C). *Right*: We repeated this procedure for different percentages (10%–100%) and found that our results can be reproduced with as little as 30% of the dataset. The asymptotic thresholds (evaluated at a population size of $N$ = 1024 neurons) do not change much beyond this point (shaded region). **(B)** Thresholds implied by the LI model obtained by extrapolating 50% of the dataset. Once again, this was similar to results obtained using the full dataset (Fig 6B).
(PDF)

**S15 Fig. Inferred readout strategy is robust to the degree of inactivation.** We extended our model to include two additional parameters $\rho_x$ and $\rho_y$ that denote fractions of neurons

inactivated in populations *x* and *y*, and derived theoretical results that account for partial inactivation of the two populations (S7 Text). We used those results to model partial inactivation of the MSTd and VIP in our dataset, and computed parameter ranges in the ($\rho_M$,$\rho_V$) parameter space (shaded areas) that are consistent with 95% confidence intervals around experimental data. (**A**) *Extensive information model*. Since an empirical trend between neural tuning and noise covariance was used to determine the structure of noise correlations, the readout weights could be uniquely determined from the observed pattern of choice correlations (CCs) independent of the extent of inactivation. Therefore the inferred readout weights remained the same as for the model that assumed complete inactivation (inferred MSTd weight scaling $a_M$ = 0.44; optimal MSTd weight scaling $a_M$ = 0.74). Nonetheless, the predictions for behavioural thresholds following inactivation of MSTd or VIP (shown in Fig 5B) are quantitatively consistent with the experimental observations (Fig 2B) only for a specific range of inactivation fractions (grey region). Specifically, the inferred readout weights predict that the thresholds should increase by a factor of 1.6 if MSTd was fully removed, yet the observed increase was only 1.2 ±0.1. This suggests that MSTd could neither have been completely inactivated nor remained completely intact, leading to the exclusion of the regions close to the left and right boundaries. For the EI model, therefore, partial inactivation of MSTd was a better match to the behavioural data. Similarly, inactivating about half of VIP is predicted to significantly reduce the threshold (Fig 8C – top panel). Since this was not observed experimentally, the inactivation parameters within the central horizontal band around 0.5 are excluded from the grey region that is consistent with data. Even with partial inactivation, therefore, the extensive information model implies that the brain underweights MSTd compared to optimal, just as reported in the main text where we assumed complete inactivation. (**B**) *Limited information model*. Noise correlations in the limited information model, unlike the extensive information model, were not known *a priori*, but were instead fit to explicitly account for the behavioural effects of inactivation. Consequently, both the readout weights and the inactivation fractions are jointly constrained by the behavioural thresholds observed after inactivating these brain areas. Thus the set of inactivation fractions consistent with data co-varied with readout weights. Shaded regions represent fraction of cortex inactivated for MSTd and VIP that were consistent with observed behavioural thresholds following inactivation (within 95% confidence intervals) assuming three different values of the scaling of MSTd readout weights ($a_M$ = 0.95, 0.85, and 0.75, shown in red, green, and blue). The solution space that was consistent with our data (shaded areas) contracted as the scaling of MSTd weights decreased, with no solutions for $a_M < 0.74$. In contrast to the extensive information model, the limited information model attributes experimental results to overweighting MSTd compared to optimal decoding in all cases (which would have $a_M$ within the intervals [0.87 0.93], [0.75 0.81], and [0.6 0.64] respectively, again to remain consistent with 95% confidence intervals of behavioural thresholds), just as we reported in the main text assuming complete inactivation. Thus the qualitative behaviour of the limited information model was robust to incomplete inactivation by Muscimol.
(PDF)

**S16 Fig. Recurrent neural network.** We extended our model to incorporate recurrent connections and derived theoretical results relating the connectivity matrix to the behavioural and neuronal effects of inactivation in steady-state (S8.1 Text). Recall that decoding weights were inferred in the subspace of the leading eigenmodes of the response covariance. Therefore, it is clear that our main results will not be affected by recurrent weights that do not significantly alter neural response along the principal components of covariance in MSTd (*M*) and VIP (*V*). Instead, we constructed a specific recurrent scheme that would couple responses along the leading modes (S8.2 Text), and used our theoretical results to test whether there exist

connection strengths ($c$) that leave our main conclusions unaltered. **(A)** Schematic of a recurrent neural network comprising the two brain areas – MSTd ($M$) and VIP ($V$). **(B)** Recurrent connectivity matrices for the extensive (EI) and limited information (LI) models. **(C)** Unlike the purely feedforward model, slopes of the tuning curves of individual neurons in this recurrent network are altered when one of the two brain areas is inactivated. **(D)** Ratio of thresholds after inactivating one of the areas to the behavioural threshold observed in the intact brain, as a function of the overall connection strength ($c$) between the areas. For appropriate choice of connection strengths (dotted line), the behavioural effects of inactivation are consistent with the experimentally observed outcomes, and nearly identical to the feedforward network for both limited and extensive information models.
(PDF)

**S1 Table. Model parameters and predictions for visual condition.** Model parameters and predicted changes in CCs following inactivation for the two covariance models, shown as median ± central quartile range. ($^{\dagger}$Values correspond to when decoder is inferred using a rank-two approximation of the covariance.). See Table 1 in main text for vestibular condition.
(PDF)

**S1 Text. Choice probability and choice correlation.**
(PDF)

**S2 Text. Optimal thresholds and coarse-grained covariance.**
(PDF)

**S3 Text. Effects of suboptimal decoding on behavioural threshold.**
(PDF)

**S4 Text. Effect of suboptimal decoding on choice correlations.**
(PDF)

**S5 Text. Combining choice correlations and inactivation effects.**
(PDF)

**S6 Text. Effect of measurement uncertainty.**
(PDF)

**S7 Text. Modeling partial inactivation.**
(PDF)

**S8 Text. Recurrent network model.**
(PDF)

**S9 Text. Effect of selective inactivation on choice correlations in the non-inactivated area.**
(PDF)

## Acknowledgments

We thank Adam Zaidel, Yong Gu, & Aihua Chen for performing the neural recordings, as well as Sheng Liu & Yong Gu for performing the muscimol inactivation experiments.

## Author Contributions

**Conceptualization:** Kaushik J. Lakshminarasimhan, Alexandre Pouget, Gregory C. DeAngelis, Xaq Pitkow.

**Data curation:** Kaushik J. Lakshminarasimhan.

**Formal analysis:** Kaushik J. Lakshminarasimhan, Xaq Pitkow.

**Funding acquisition:** Dora E. Angelaki, Xaq Pitkow.

**Investigation:** Kaushik J. Lakshminarasimhan, Alexandre Pouget, Gregory C. DeAngelis, Dora E. Angelaki, Xaq Pitkow.

**Methodology:** Kaushik J. Lakshminarasimhan, Alexandre Pouget, Gregory C. DeAngelis, Dora E. Angelaki, Xaq Pitkow.

**Project administration:** Dora E. Angelaki, Xaq Pitkow.

**Resources:** Dora E. Angelaki.

**Supervision:** Dora E. Angelaki, Xaq Pitkow.

**Visualization:** Kaushik J. Lakshminarasimhan, Dora E. Angelaki, Xaq Pitkow.

**Writing – original draft:** Kaushik J. Lakshminarasimhan, Xaq Pitkow.

**Writing – review & editing:** Kaushik J. Lakshminarasimhan, Alexandre Pouget, Gregory C. DeAngelis, Dora E. Angelaki, Xaq Pitkow.

## References


1. Nienborg H, R. Cohen M, Cumming BG. Decision-Related Activity in Sensory Neurons: Correlations Among Neurons and with Behavior. Annual Review of Neuroscience. 2012. pp. 463–483. https://doi.org/10.1146/annurev-neuro-062111-150403 PMID: 22483043
2. Georgopoulos AP, Schwartz AB, Kettner RE. Neuronal population coding of movement direction. Science. 1986; 233: 1416–1419. https://doi.org/10.1126/science.3749885 PMID: 3749885
3. Paradiso MA. A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. Biol Cybern. 1988; 58: 35–49. PMID: 3345319
4. Pouget A, Thorpe SJ. Connectionist Models of Orientation Identification. Connection Science. 1991. pp. 127–142.
5. Seung HS, Sompolinsky H. Simple models for reading neuronal population codes. Proc Natl Acad Sci U S A. 1993; 90: 10749–53. https://doi.org/10.1073/pnas.90.22.10749 PMID: 8248166
6. Shadlen MN, Britten KH, Newsome WT, Movshon JA. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. J Neurosci. 1996; 16: 1486–1510. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8778300%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/8778300 PMID: 8778300
7. Oram MW, Földiák P, Perrett DI, Sengpiel F. The "Ideal Homunculus": decoding neural population signals. Trends Neurosci. 1998; 21: 259–265. https://doi.org/10.1016/S0166-2236(97)01216-2 PMID: 9641539
8. Chen Y, Geisler WS, Seidemann E. Optimal decoding of correlated neural population responses in the primate visual cortex. Nat Neurosci. 2006; 9: 1412–1420. https://doi.org/10.1038/nn1792 PMID: 17057706
9. Cohen MR, Newsome WT. Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. J Neurosci. 2009; 29: 6635–6648. https://doi.org/10.1523/JNEUROSCI.5179-08.2009 PMID: 19458234
10. Graf ABA, Kohn A, Jazayeri M, Movshon JA. Decoding the activity of neuronal populations in macaque primary visual cortex. Nat Neurosci. 2011; 14: 239–245. https://doi.org/10.1038/nn.2733 PMID: 21217762
11. Berens P, Ecker AS, Cotton RJ, Ma WJ, Bethge M, Tolias AS. A Fast and Simple Population Code for Orientation in Primate V1. Journal of Neuroscience. 2012. pp. 10618–10626. https://doi.org/10.1523/JNEUROSCI.1335-12.2012 PMID: 22855811
12. Gu Y, Angelaki DE, DeAngelis GC. Contribution of correlated noise and selective decoding to choice probability measurements in extrastriate visual cortex. Elife. eLife Sciences Publications Ltd; 2014;
13. Crapse TB, Basso MA. Insights into Decision-Making Using Choice Probability. J Neurophysiol. 2015; jn.00335.2015. https://doi.org/10.1152/jn.00335.2015 PMID: 26378203

14. Haefner RM, Gerwinn S, Macke JH, Bethge M. Inferring decoding strategies from choice probabilities in the presence of correlated variability. Nat Neurosci. 2013; 16: 235–42. https://doi.org/10.1038/nn.3309 PMID: 23313912

15. Pitkow X, Liu S, Angelaki DE, DeAngelis GC, Pouget A. How Can Single Sensory Neurons Predict Behavior? Neuron. Elsevier Inc.; 2015; 87: 411–423. Available: http://linkinghub.elsevier.com/retrieve/pii/S0896627315005966

16. Hanks TD, Kopec CD, Brunton BW, Duan CA, Erlich JC, Brody CD. Distinct relationships of parietal and prefrontal cortices to evidence accumulation. Nature. 2015; 520: 220–3. Available: http://www.ncbi.nlm.nih.gov/pubmed/25600270 https://doi.org/10.1038/nature14066 PMID: 25600270

17. Raposo D, Kaufman MT, Churchland AK. A category-free neural population supports evolving demands during decision-making. Nat Neurosci. 2014; 17: 1784–1792. https://doi.org/10.1038/nn.3865 PMID: 25383902

18. Chen A, Gu Y, Liu S, Deangelis GC, Angelaki DE. Evidence for a causal contribution of macaque vestibular, but not intraparietal, cortex to heading perception. J Neurosci. 2016;

19. Katz L, Yates J, Pillow JW, Huk AC. Dissociated functional significance of decision-related activity in the primate dorsal stream. Nature. 2016; 535: 285–288. https://doi.org/10.1038/nature18617 PMID: 27376476

20. Moreno-Bote R. B. J, Kanitscheider I, Pitkow X, Latham PE, Pouget A. Information-limiting correlations. Nat Neurosci. 2014; 17: 1410–1417. https://doi.org/10.1038/nn.3807 PMID: 25195105

21. Chen A, Deangelis GC, Angelaki DE. Functional specializations of the ventral intraparietal area for multisensory heading discrimination. J Neurosci. 2013; 33: 3567–81. https://doi.org/10.1523/JNEUROSCI.4522-12.2013 PMID: 23426684

22. Gu Y, DeAngelis GC, Angelaki DE. Causal Links between Dorsal Medial Superior Temporal Area Neurons and Multisensory Heading Perception. Journal of Neuroscience. 2012. pp. 2299–2313. https://doi.org/10.1523/JNEUROSCI.5154-11.2012 PMID: 22396405

23. Gu Y, Angelaki DE, Deangelis GC. Neural correlates of multisensory cue integration in macaque MSTd. Nat Neurosci. 2008; 11: 1201–10. https://doi.org/10.1038/nn.2191 PMID: 18776893

24. Britten KH, Newsome WT, Shadlen MN, Celebrini S, Movshon JA. A relationship between behavioral choice and the visual responses of neurons in macaque MT. Vis Neurosci. 1996; 13: 87–100. PMID: 8730992

25. Zohary E, Shadlen MN, Newsome WT. Correlated neuronal discharge rate and its implications for psychophysical performance. Nature. 1994; 370: 140–143. https://doi.org/10.1038/370140a0 PMID: 8022482

26. Abbott LF, Dayan P. The effect of correlated variability on the accuracy of a population code. Neural Comput. 1999; 11: 91–101. PMID: 9950724

27. Sompolinsky H, Yoon H, Kang K, Shamir M. Population coding in neuronal systems with correlated noise. Physical Review E. 2001. https://doi.org/10.1103/PhysRevE.64.051904 PMID: 11735965

28. Averbeck BB, Lee D. Effects of noise correlations on information encoding and decoding. J Neurophysiol. 2006; 95: 3633–3644. https://doi.org/10.1152/jn.00919.2005 PMID: 16554512

29. Gu Y, Liu S, Fetsch CR, Yang Y, Fok S, Sunkara A, et al. Perceptual learning reduces interneuronal correlations in macaque visual cortex. Neuron. 2011; 71: 750–761. https://doi.org/10.1016/j.neuron.2011.06.015 PMID: 21867889

30. Liu S, Gu Y, DeAngelis GC, Angelaki DE. Choice-related activity and correlated noise in subcortical vestibular neurons. Nat Neurosci. 2013; 16: 89–97. https://doi.org/10.1038/nn.3267 PMID: 23178975

31. Wohrer A, Romo R, Machens C. Linear readout from a neural population with partial correlation data. Adv Neural Inf Process Syst 23. 2010; 2469–2477.

32. Wohrer A, Machens CK. On the Number of Neurons and Time Scale of Integration Underlying the Formation of Percepts in the Brain. PLoS Comput Biol. 2015; 11: 1–38. https://doi.org/10.1371/journal.pcbi.1004082 PMID: 25793393

33. Shamir M, Sompolinsky H. Implications of neuronal diversity on population coding. Neural computation. 2006. pp. 1951–1986. https://doi.org/10.1162/neco.2006.18.8.1951 PMID: 16771659

34. Ecker AS, Berens P, Tolias AS, Bethge M. The Effect of Noise Correlations in Populations of Diversely Tuned Neurons. Journal of Neuroscience. 2011. pp. 14272–14283. https://doi.org/10.1523/JNEUROSCI.2539-11.2011 PMID: 21976512

35. Hu Y, Zylberberg J, Shea-Brown E. The Sign Rule and Beyond: Boundary Effects, Flexibility, and Noise Correlations in Neural Population Codes. PLoS Comput Biol. Public Library of Science; 2014; 10.

36. Haefner RM, Berkes P, Fiser J. Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. Neuron. 2016; https://doi.org/10.1016/j.neuron.2016.03.020 PMID: 27146267

37. Averbeck BB, Latham PE, Pouget A. Neural correlations, population coding and computation. Nat Rev Neurosci. 2006; 7: 358–366. https://doi.org/10.1038/nrn1888 PMID: 16760916

38. Schneidman E, Bialek W, II MJB. Synergy, Redundancy, and Independence in Population Codes. J Neurosci. 2003; 23: 11539–11553. PMID: 14684857

39. Beck JM, Ma WJ, Pitkow X, Latham PE, Pouget A. Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. Neuron. 2012. pp. 30–39. https://doi.org/10.1016/j.neuron.2012.03.016 PMID: 22500627

40. Schoups AA, Vogels R, Orban GA. Human perceptual learning in identifying the oblique orientation: retinotopy, orientation specificity and monocularity. J Physiol. 1995; 483: 797–810. PMID: 7776259

41. Jehee JFM, Ling S, Swisher JD, van Bergen RS, Tong F. Perceptual learning selectively refines orientation representations in early visual cortex. J Neurosci. 2012; 32: 16747–53a. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3575550&tool=pmcentrez&rendertype=abstract https://doi.org/10.1523/JNEUROSCI.6112-11.2012 PMID: 23175828

42. Li W, Piëch V, Gilbert CD. Perceptual learning and top-down influences in primary visual cortex. Nat Neurosci. 2004; 7: 651–657. https://doi.org/10.1038/nn1255 PMID: 15156149

43. Kopec CD, Erlich JC, Brunton BW, Deisseroth K, Brody CD. Cortical and Subcortical Contributions to Short-Term Memory for Orienting Movements. Neuron. Cell Press; 2015; 88: 367–377. https://doi.org/10.1016/j.neuron.2015.08.033 PMID: 26439529

44. Wong K-F, Wang X-J. A recurrent network mechanism of time integration in perceptual decisions. J Neurosci. 2006; 26: 1314–28. Available: http://www.jneurosci.org/content/26/4/1314.full https://doi.org/10.1523/JNEUROSCI.3733-05.2006 PMID: 16436619

45. Machens CK, Romo R, Brody CD. Flexible Control of Mutual Inhibition: A Neural Model of Two-Interval Discrimination. Science (80-). AAAS; 2005; 307: 1121–4. https://doi.org/10.1126/science.1104171 PMID: 15718474

46. Park IM, Meister MLR, Huk AC, Pillow JW. Encoding and decoding in parietal cortex during sensorimotor decision-making. Nat Neurosci. 2014; 17: 1395–1403. Available: https://doi.org/10.1038/nn.3800 PMID: 25174005

47. Nienborg H, Cumming BG. Decision-related activity in sensory neurons reflects more than a neuron's causal effect. Nature. 2009; 459: 89–92. https://doi.org/10.1038/nature07821 PMID: 19270683

48. de Lafuente V, Jazayeri M, Shadlen MN. Representation of accumulating evidence for a decision in two parietal areas. J Neurosci. 2015; 35: 4306–18. Available: http://www.jneurosci.org/content/35/10/4306.full https://doi.org/10.1523/JNEUROSCI.2451-14.2015 PMID: 25762677

49. Yang H, Kwon SE, Severson KS, O'Connor DH. Origins of choice-related activity in mouse somatosensory cortex. Nat Neurosci. 2015; 19: 127–134. https://doi.org/10.1038/nn.4183 PMID: 26642088

50. Wimmer K, Compte A, Roxin A, Peixoto D, Renart A, de la Rocha J. Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. Nat Commun. Nature Publishing Group; 2015; 6: 6177. https://doi.org/10.1038/ncomms7177 PMID: 25649611

51. Gao P, Ganguli S. On simplicity and complexity in the brave new world of large-scale neuroscience. Current Opinion in Neurobiology. Elsevier Ltd; 2015. pp. 148–155.

52. Beck J, Pouget A. Insights from a Simple Expression for Linear Fisher Information in a Recurrently Connected Population of Spiking Neurons. Neural Comput. 2011; 23: 1484–1502. https://doi.org/10.1162/NECO_a_00125 PMID: 21395435

53. Britten KH, Shadlen MN, Newsome WT, Movshon JA. The analysis of visual motion: a comparison of neuronal and psychophysical performance. J Neurosci. 1992; 12: 4745–4765. https://doi.org/10.1.1.123.9899 PMID: 1464765

54. Green DM, Swets JA. Signal detection theory and psychophysics [Internet]. New York Wiley. Wiley; 1966. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Signal+detection+theory+and+psychophysics#0

55. Kang I, Maunsell JHR. Potential confounds in estimating trial-to-trial correlations between neuronal response and behavior using choice probabilities. Journal of Neurophysiology. 2012.