bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Inferring decoding strategies for multiple correlated neural populations

Kaushik J Lakshminarasimhan[1], Alexandre Pouget[3,4], Gregory C DeAngelis[4], Dora E Angelaki[1,5,#], Xaq Pitkow[1,5,#,*]

[1]Department of Neuroscience, Baylor College of Medicine, Houston, USA

[3]Department of Basic Neuroscience, University of Geneva, Switzerland

[4]Department of Brain and Cognitive Sciences, University of Rochester, Rochester, USA

[5]Department of Electrical and Computer Engineering, Rice University, Houston, USA

[#]These authors contributed equally.

[*]Corresponding author: xaq@rice.edu

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 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 brain area can predict an animal's 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 can be distinguished by an experiment that measures correlations while inactivating different brain areas.

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Although much is known about how single neurons encode information about stimuli, how neurons contribute to 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

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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. 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

Decoding framework

We consider a linear feedforward network in which the firing rates $\mathbf{r}$ of the neurons are combined linearly using weights $\mathbf{w}$ to yield a locally unbiased estimate $\hat{s}$ of the stimulus according to $\hat{s} = \mathbf{w}^T(\mathbf{r} - \mathbf{f}(s_0))$, where $\mathbf{f}(s_0)$ is the mean response to a reference stimulus s_0. In each trial, the animal is assumed to reach a binary decision given by $\mathrm{sgn}(\hat{s}) = \pm 1$, where sgn is the signum function. For a decoder that linearly reads out neurons from two subpopulations, x and y, the estimate $\hat{s}$ can be expressed as:

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

$$\hat{s} = a_x \hat{s}_x + a_y \hat{s}_y \quad (1)$$

where $\hat{s}_x = \mathbf{w}_x^T(\mathbf{r}_x - \mathbf{f}_x(s_0))$ and $\hat{s}_y = \mathbf{w}_y^T(\mathbf{r}_y - \mathbf{f}_y(s_0))$ denote unbiased estimates derived from neurons in subpopulations x and y respectively. 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 both populations together using $\mathbf{w} = [a_x\mathbf{w}_x \quad a_y\mathbf{w}_y]$. The form of equation (1) has two advantages: (*i*) it is easy to identify and compare the relative contributions of the two areas to behaviour through the ratio a_x/a_y, and (*ii*) one can dissociate how the weight *patterns* ($\mathbf{w}_x$ and $\mathbf{w}_y$) and their *scales* (a_x and a_y) 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 discrimination thresholds before (ϑ) and after (ϑ_{-x} and ϑ_{-y}) inactivating population x or y (**Figure 1b**).

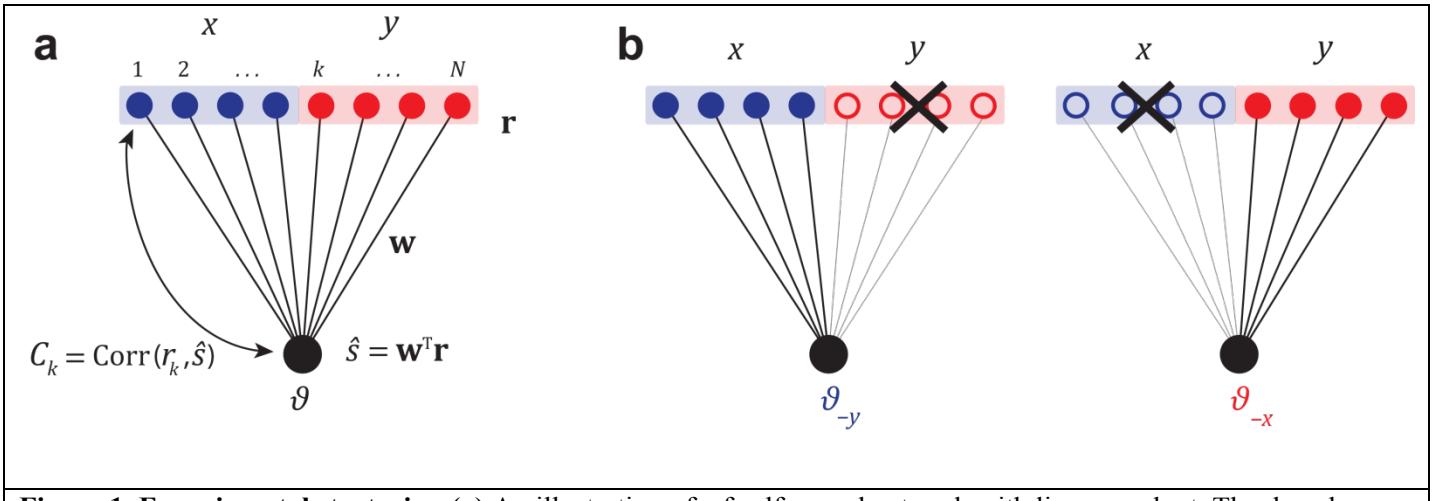

Figure 1. Experimental strategies. (a) An illustration of a feedforward network with linear readout. The decoder linearly combines the activity **r** of neurons in populations x and y with weights **w**, to produce an estimate $\hat{s}$ of the stimulus. Activity of individual neurons r_k is correlated 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 **equation 2.1**. **(b)** In inactivation experiments, the neurons from each population are inactivated and the resulting changes in behavioural threshold are recorded.

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We would like to use these experimental measurements to identify the relative behavioural contributions of two brain areas. Therefore we will present a technique to infer neuronal weights in two brain areas, focusing primarily on how to extract the scaling factors, a_x and a_y, of the brain areas rather than the fine structure, $\mathbf{w}_x$ and $\mathbf{w}_y$, of the decoding weights. We first present some results that allow us to examine the pattern of choice correlations of neurons in both areas to characterize the degree of suboptimality in decoding. We will then show how to combine choice correlations with inactivation results to obtain quantitative estimates of the relative scaling of readout weights in those areas.

Analysis of choice correlations

Choice correlation of a neuron *k* is the correlation coefficient, across repeated trials with the same stimulus *s*, between its response r_k and the animal's estimate of the stimulus $\hat{s}$, $C_k = \mathrm{Corr}(\hat{s}, r_k|s)$. It has recently been shown that readout weights are optimal only if neuronal choice correlations all satisfy the following relation[15] (**Supplementary note S1**):

$$C_{k,\mathrm{opt}} = \frac{\vartheta}{\vartheta_k} \tag{2.1}$$

where $C_{k,\mathrm{opt}}$ is the choice correlation of neuron k expected from optimal decoding, ϑ_k is the discrimination threshold of neuron *k*, and ϑ is the behavioural discrimination threshold. Therefore if neurons from both areas satisfy the above equation, this gives us strong evidence that the neuronal weights and consequently their relative scales $\boldsymbol{a} = (a_x, a_y)$ are optimal. As we will see later, the exact values of $\boldsymbol{a}$ can then be directly extracted from the behavioural thresholds ϑ_{-x} and ϑ_{-y} following inactivation of those areas.

The pattern of choice correlations generated by any generic suboptimal decoder is more complicated, as it depends explicitly on the structure of noise covariance[14]. For a population of *N* neurons, the covariance Σ describes the noise power along *N* orthogonal noise modes. Each of these modes contributes to the overall choice correlation according to (**Supplementary note S2**):

$$C_k = \sum_{i=1}^{N} \beta_i C_{k,\mathrm{opt}}^i \tag{2.2}$$

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In this expression we have decomposed the optimal pattern of choice correlations $C_{k,\mathrm{opt}}$ into components $C^i_{k,\mathrm{opt}}$ originating from the different noise modes of Σ, with $\sum_{i=1}^{N} C^i_{k,\mathrm{opt}} = C_{k,\mathrm{opt}}$. The multipliers β_i reflect the extent of suboptimality. When decoding weights are optimal, every multiplier $\beta_i = 1$, so the above equation reduces to **equation 2.1**.

In principle, it is very difficult to estimate all of the multipliers β_i because the components $C^i_{k,\mathrm{opt}}$ depend on the individual noise modes of Σ (**Methods M1 – equation 4**). Directly measuring Σ 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 are carried out, it would be impossible to get an accurate assessment of the fine structure of covariance with limited data due to errors arising from finite measurement density[24]. Fortunately, since neuronal choice correlations are measurably large, it follows that one can infer 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] (**Figure S1**). It is possible to determine properties of the leading modes of covariance without large-scale recordings, and we will consider two ways producing two different noise models: *extensive information* and *limited information*.

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] (**Methods M8 — equation 7.1**). Although the resulting covariance matrix is unlikely to

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

capture fine details accurately, if the model is reasonable then most of the variance would be captured by the leading modes.

When decoding two populations *x* and *y*, one has to consider at least two leading modes to capture the two underlying degrees of freedom decoded by scaling factors a_x and a_y. In this minimal case, choice correlations are given by $C_k = \beta_1 C^1_{k,\mathrm{opt}} + \beta_2 C^2_{k,\mathrm{opt}}$. We can compute $C^1_{k,\mathrm{opt}}$ and $C^2_{k,\mathrm{opt}}$ from the leading modes of covariance (**Methods M1** – **equation 4**), and use them to estimate β_1 and β_2 by linear regression. If there are two dominant noise modes and they affect both populations, then we can approximate Σ with a rank-two noise covariance matrix composed of both independent (ε_{xx} and ε_{yy}) and correlated (ε_{xy}) noise between the two areas (**Supplementary note S3**). If the two modes were actually uncorrelated, with $\varepsilon_{xy} = 0$, so that each mode affects just one population, then the multipliers β_1 and β_2 would be specific to neurons in each population and therefore correspond to β_x and β_y.

A characteristic feature of extensive information models is that the amount of information in the neural activity is very large because it grows with population size[33–35], hence the name. The amount of information extracted by a decoder restricted to the subspace spanned by the few dominant components of covariance cannot be greater than the information available in that subspace. For a model with extensive information, this subset would be a tiny fraction of the total information available in the population. Although this restriction is justified by the large magnitude of neuronal choice correlations, the choice of this noise model is only justified under the assumption that the brain is radically suboptimal.

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 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

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 $\mathbf{f}(\mathrm{s} + \delta\mathrm{s}) - \mathbf{f}(\mathrm{s}) \approx \delta s\, \mathbf{f}'(s)$. 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 ε, represents the variance of information-limiting correlations. With increasing population size, both the signal and this noise component grow identically, resulting in no further improvement in signal-to-noise ratio, and thus no improvement in discriminability. In general, ε 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 $1/\varepsilon$.

When dealing with two populations x and y, 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 two populations to receive both distinct and shared information. The covariance between two neurons in this more general 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 pair of neurons are both from the same population x (ε_{xx}), both from y (ε_{yy}), or from different populations (ε_{xy}) (**Methods M9 – equation 8**). For a large population with this noise structure, the total information content within the x and y subpopulations alone are by construction equal to $1/\varepsilon_{xx}$ and $1/\varepsilon_{yy}$ respectively. The information in both populations together is limited as well, once again by the $\mathbf{f}'\mathbf{f}'^{\mathrm{T}}$ component of the covariance. Depending on ε_{xy}, the two subpopulations may contain completely redundant, independent, or synergistic information[36,37]. In case the two populations receive information from the same source, then $\varepsilon_{xx} = \varepsilon_{yy} = \varepsilon_{xy}$ yielding the familiar form of information-limiting correlations[15,20] $\Sigma_{\mathrm{IL}} = \Sigma + \varepsilon\mathbf{f}'\mathbf{f}'^{\mathrm{T}}$.

Correlations that limit information within a single neural population introduce redundancy. As a consequence, many different decoding weights can extract essentially the same information. The system is

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

then robust to some suboptimal decoding, which makes it easier to achieve near-optimal behavioural performance[15]. In the noise model for two populations described above, this is also true for each population individually. We can generalize this robustness in our framework by considering separate decoders of each population that produce estimates, $\hat{s}_x$ and $\hat{s}_y$, that are near-optimal for their 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 **equation 1**. While information-limiting correlations within each area would make the system generally robust to the choice of weight patterns $\mathbf{w}_x$ or $\mathbf{w}_y$, suboptimality could yet arise from an incorrect scaling (a_x and a_y) of the individual near-optimal estimates. This is because after the dimensionality reduction from large redundant populations down to single unbiased estimates per population, there is no redundancy left: just one degree of freedom remains for the decoder, so different ways of combining the estimates are not equivalent.

If the brain indeed combines activity from different areas suboptimally in this manner, then simplifying **equation 2.2** in the presence of information-limiting correlations gives choice correlations within each area that are not equal to the optimal choice correlations, but are proportional to them (**Supplementary note S5**):

$$C_k = \beta \frac{\vartheta}{\vartheta_k} \tag{2.3}$$

Under these conditions, choice correlations in different areas *x* and *y* may have different multipliers β, say β_x and β_y, which depend on the scaling of the two brain areas and on the covariance between the two estimates derived from them. These multipliers β_x and β_y can be directly identified by regressing measured choice correlations against ϑ/ϑ_k, the choice correlations predicted for optimal decoding.

Combining choice correlations and inactivation effects to infer decoding weights

In the previous section, we showed how to reduce the fine structure of choice correlations down to one number for each population — β_x and β_y. We will now show how these multipliers can be used, together with the behavioural thresholds ϑ_{-x} and ϑ_{-y} following inactivation of areas *x* and *y*, respectively, to infer the relative scaling of their weights a_x and a_y. Inactivating an area is equivalent to setting the scaling of

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

weights in that area to zero, so from **equation 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 $\vartheta_{-x}^2 \approx \varepsilon_{yy}$ and $\vartheta_{-y}^2 \approx \varepsilon_{xx}$. If populations *x* and *y* are uncorrelated ($\varepsilon_{xy} = 0$), then the ratio of weight scalings can be factorized into a product of ratios (**Supplementary note S6**):

$$\frac{a_x}{a_y} = \frac{\beta_x}{\beta_y}\frac{\varepsilon_{yy}}{\varepsilon_{xx}} \approx \frac{\beta_x}{\beta_y}\frac{\vartheta_{-x}^2}{\vartheta_{-y}^2} \tag{3.1}$$

where the two independent factors represent outcomes of correlational and causal studies. If readout is optimal, then the multipliers β_x and β_y are both equal to one, so $a_x/a_y = \vartheta_{-x}^2/\vartheta_{-y}^2$. 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 roles of the brain areas (**Figure S2**). 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, **equation 3.1** 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 ε_{xy} and the ratio of scales no longer factorizes:

$$\frac{a_x}{a_y} \approx \left(\frac{\beta_x}{\beta_y}\frac{\vartheta_{-x}^2}{\vartheta_{-y}^2} - \gamma\right)\left(1 - \frac{\beta_x}{\beta_y}\gamma\right)^{-1} \tag{3.2}$$

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 **equations 3.1** and **3.2** to compute the optimal weight scaling factors simply by setting both β_x and β_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.

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Figure 2a**) using vestibular (inertial motion) and/or visual (optic flow) cues (**Methods M4**; see also refs. [21,23]). At the end of each trial, the animal reported whether their perceived heading $\hat{s}$ was leftward ($\hat{s} < 0°$) or rightward ($\hat{s} > 0°$) 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 (**Methods M5**). 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 (**Figure 2b –** left) for both the vestibular and visual conditions. This difference in CC between areas could not be attributed to differences in neuronal thresholds ϑ_k (**Figure 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 M6**; **Figure S3**). 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] (**Figure 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] (**Figure 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 (**Figure S4**).

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

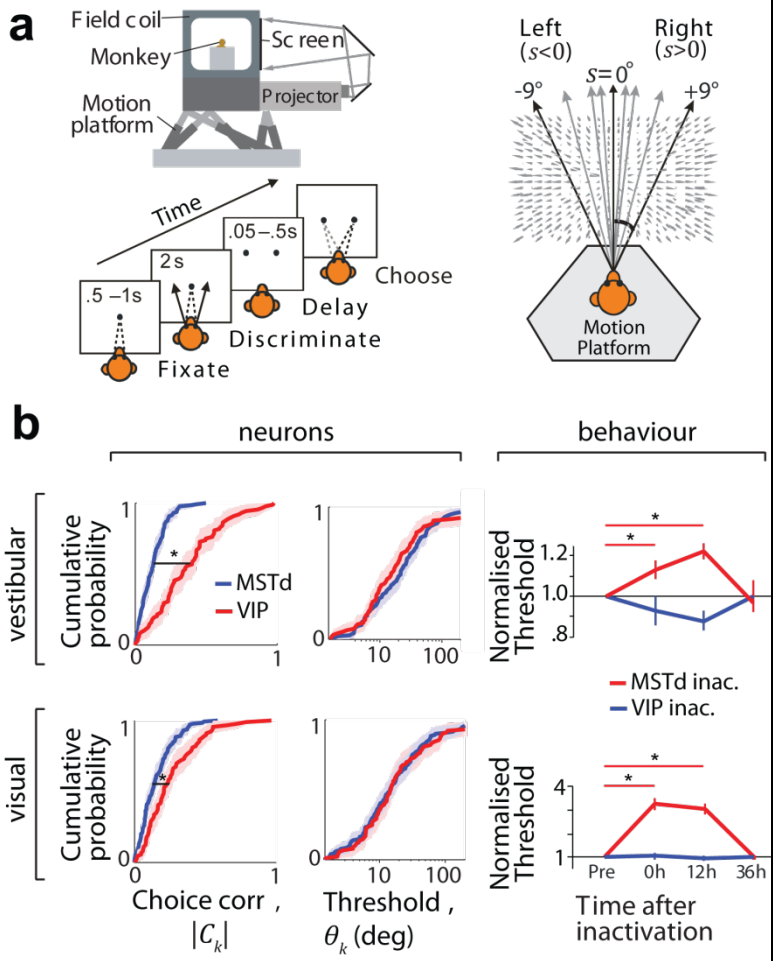

Figure 2. Choice-related activity and effects of inactivation. (a) Behavioural task: the monkey sits on a motion platform facing a screen. He fixates on a small target at the center of the screen, and then we induce a self-motion percept by moving the platform (vestibular condition) or by displaying 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 and MSTd. 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].

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 **equation 2.1**. 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 $\hat{s}$; they only track the animal's binary choice. Instead they measure a related quantity known as choice probability defined as the probability

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 probabilities to choice correlations using a known relation[14] before further analyses (**Methods M7**). 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 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 **equation 2.1** to generate predictions for optimal CCs that accounted for our convention (**Methods M7**).

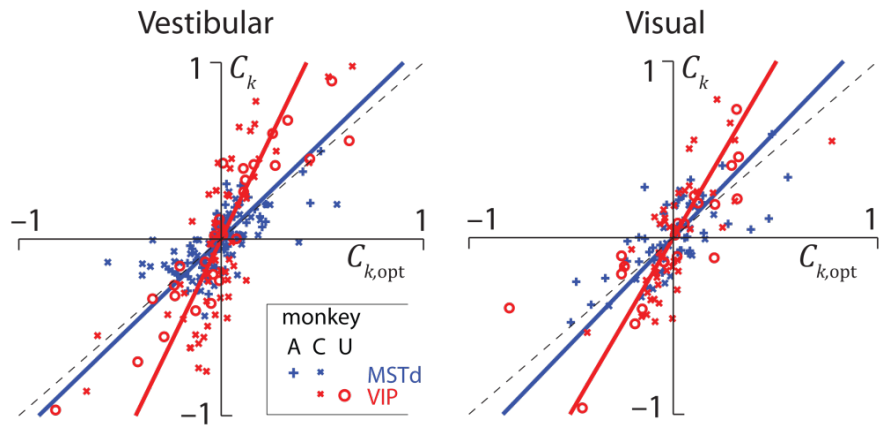

Figure 3. Readout is not optimal. Whereas the experimentally measured choice correlations (C_k) of neurons in MSTd (blue) for both the vestibular (left) and the visual (right) condition are well described by the optimal predictions ($C_{k,\mathrm{opt}}$), those of VIP neurons are systematically greater (red). This observation was consistent across all monkeys (see **Supplementary Figure S4a** for monkey *X*). Solid lines correspond to the best linear fit. Vestibular data replotted from Ref.[15] with different sign convention (**Methods M7**).

Figure 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. 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

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 supplementary notes. 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 (**Methods M8 – equation 7.1**). This relationship between noise and signal correlations has a substantially steeper slope in VIP than in MSTd (MSTd: m_M=0.19±0.08; VIP: m_V=0.70±0.16, **Figure S5**). 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. Since correlations between VIP and MSTd populations were not measured experimentally, we explored different correlation matrices (**Methods M8 – equation 7.2**).

In the limited information case, we added correlations that limited the total information content across the two populations (**Methods M9 – Equation 8**). For this latter case, we relied on behavioural thresholds before and after inactivation, and choice correlations, to determine the magnitudes of noise within (ε_{MM} and ε_{VV}) and between (ε_{MV}) areas (**Methods M9**). 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

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

distributing neurons equally among the two areas was made only for convenience and has no bearing on the result as explained later.

Figure 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 (**Figure 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.

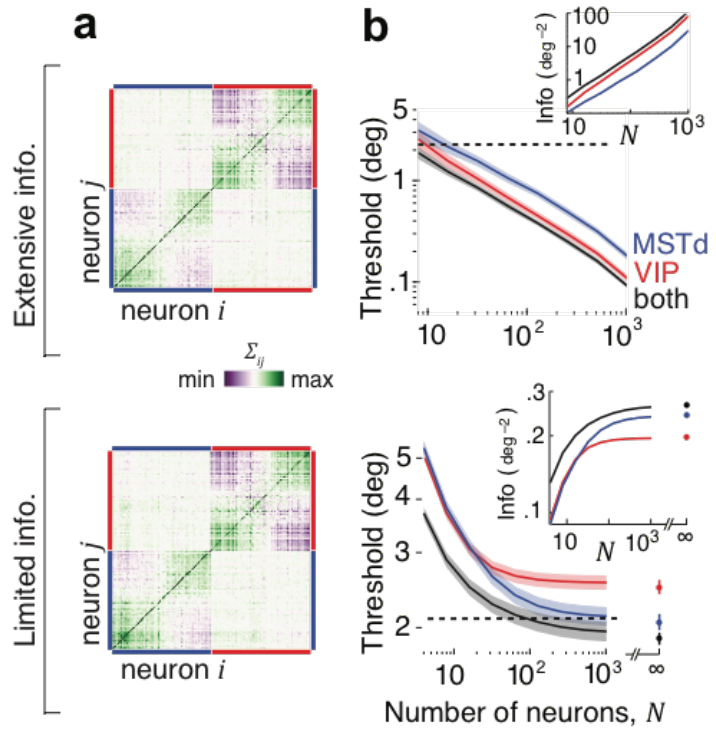

Figure 4. Covariance structure of Extensive and Limited information models. (a) Matrix of covariances Σ_{ij} among neurons in MSTd and VIP (N=128). Top: Extensive information model constructed by sampling according to the empirical relationship in **Supplementary Figure S5**, for the case when the two 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.

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Figure S6a**). 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 (**Figure S6b**). 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 (**Figure 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 (**Methods M1 – equation 4**). 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 (**Figure 5b**). However the slopes β_M and β_V were significantly different from 1 (β_M=0.73, 95% CI =[0.63 0.84]; β_V=2.38, 95% CI =[2.2 2.57]) implying that the weight scalings a_M and a_V must be suboptimal even within the two-dimensional subspace. Since we knew the magnitudes of ε_{MM} and ε_{VV} for this noise model from pairwise recordings (**Table 1**), we applied the exact rather than approximate form of **equation 3.1** 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 to zero so post-inactivation thresholds are given by the variance along the leading noise mode specific to the active area

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

($\mathbf{u}^M$ 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 large increase in threshold (**Figure 5c**, red vs black) whereas VIP inactivation is predicted to have little or no effect (**Figure 5c**, blue vs black; see **Figure S7** for visual condition). This correspondence to experimental inactivation results is remarkable because the procedure to deduce scalings 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 scalings (**Table 1**) was smaller than that produced by inferred weights (**Figure 5c**, green vs black) implying that the brain indeed weighted the two areas suboptimally.

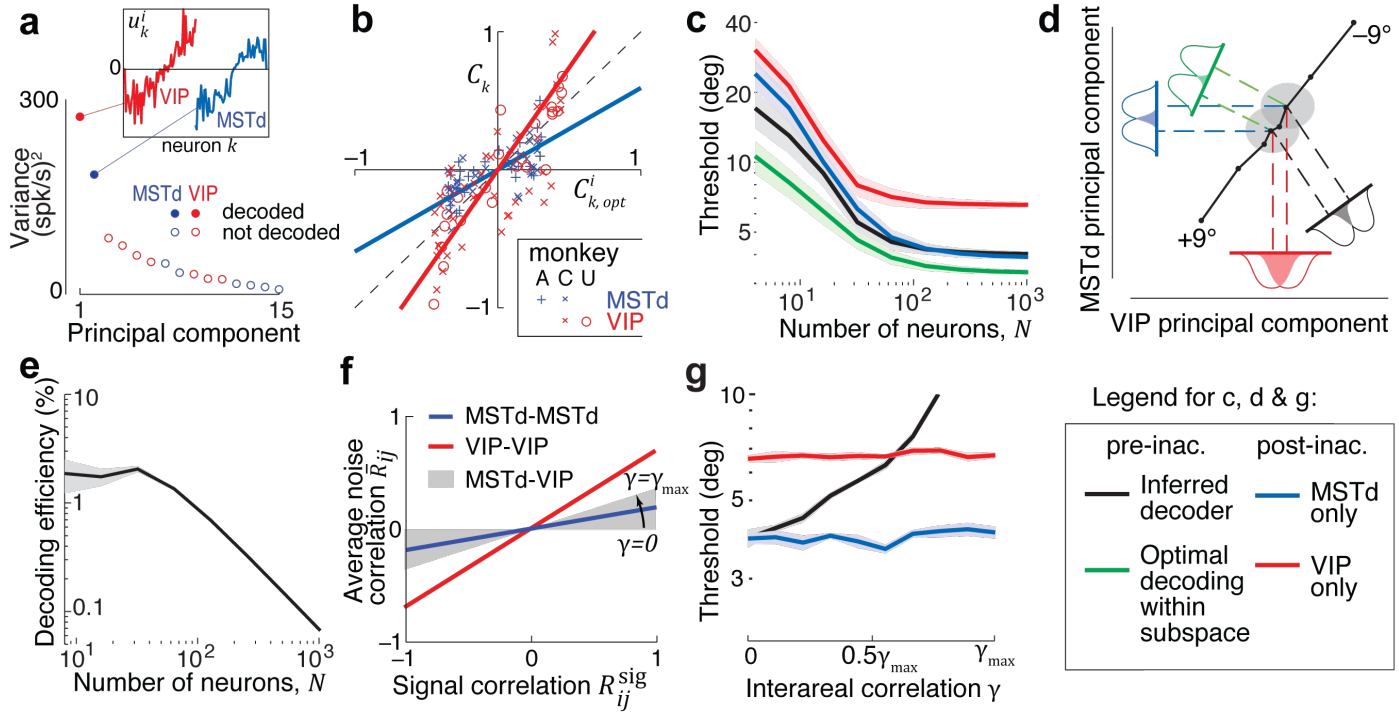

Figure 5. Decoder inferred using the extensive information model. (a) Decoding weights were inferred in the subspace of 2 leading principal components of noise covariance (solid circles). Inset: These components lie entirely within the space spanned by neurons in one of the two brain regions. Components 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 **Figure 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

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 M8 – Equation 7.2**). 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 γ, 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 γ. Shaded regions in (c), (e), and (g) represent ±1 SEM.

The above findings are explained graphically in **Figure 5d** by projecting the relevant quantities (tuning curves $\mathbf{f}(s)$, noise covariance Σ, decoding weights $\mathbf{w}$) onto the subspace of the first two principal components ($\mathbf{u}^M$ and $\mathbf{u}^V$) of the noise covariance Σ. 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

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	$\zeta_M = 2.2 \pm 0.3$ $\zeta_V = 1.3 \pm 0.1$	$\zeta_M = 0.9 \pm 0.4$ $\zeta_V = 1.3 \pm 0.4$

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.)

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 (**Figure 5c**). As illustrated in **Figure 5e**, the fraction of available information recovered by this decoder (η) 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 (**Figure S8**), 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 γ (**Figure 5f**), and used **equation 3.2** to infer weight scalings for each case. We used these scalings to generate behavioural predictions for different values of γ. Predictions for one example value of these correlations are shown in **Figure S9**. Behavioural predictions progressively worsened as a function of the strength of noise correlations between MSTd and VIP: for this model, even

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

weak but nonzero interareal correlations imply that inactivating area VIP should improve behavioural performance (**Figure 5g**).

Limited information model

In the presence of information-limiting correlations, choice correlations must be proportional to the ratio of behavioural to neuronal thresholds (**Equation 2.3**). This was indeed the case both in MSTd and VIP as we showed already in **Figure 3**. Those slopes correspond to the multipliers β_M and β_V for this model, and were found to be different for the two areas (**Table 1**).

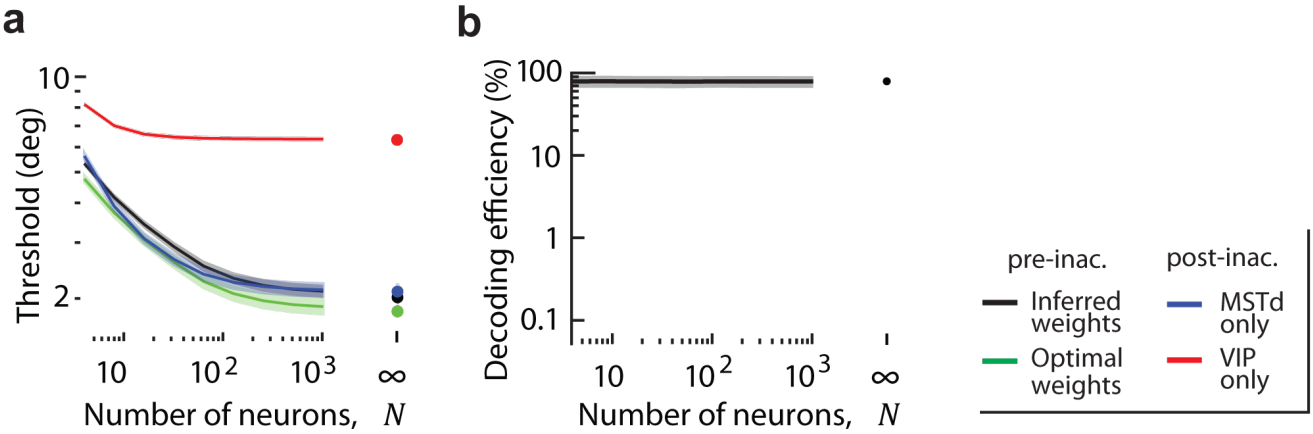

Figure 6. Decoder inferred using the limited information model. (a) Like decoding in the presence of extensive information, this decoder is suboptimal (black vs green), and can account for the behavioural effects of inactivation. **(b)** Unlike decoding in the extensive information model, the efficiency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

As we noted earlier, unlike the leading modes of noise in the extensive information model, the magnitudes of information-limiting correlations (ε_{MM}, ε_{VV}, and ε_{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 β_M and β_V derived from choice correlations as additional constraints, we can simultaneously infer the magnitude of information-limiting correlation within each area (ε_{MM} and ε_{VV}), the correlated component of the noise (ε_{MV}), and weight scalings (a_M and a_V) (**Methods M9**). 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 (**Figure 6a;** see **Figure S10** for visual condition). This was because the scaling of weights

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 η of this decoder did not depend on the size of the population being decoded (**Figure 6b**, vestibular: $\eta = 79 \pm 13\%$) because neurons in this model carry a lot of redundant information.

All analyses above were performed on neural data in the central 400ms of the trials following earlier work. However our conclusions are robust to the specific time (**Figure S11**) and duration (**Figure S12**) of the analysis window. Additionally, 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 (**Figure S13**). 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 (**Supplementary note S8**; **Figure S14**). 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 if the above assumption is compromised by recurrent connections between MSTd and VIP (**Supplementary note S9**; **Figure S15**).

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, Figure 7a**). 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 (**Figure 5g** and **Figure S9**), and the neuronal weights in

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Figure 7b – 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 (**Figure 7b – bottom**).

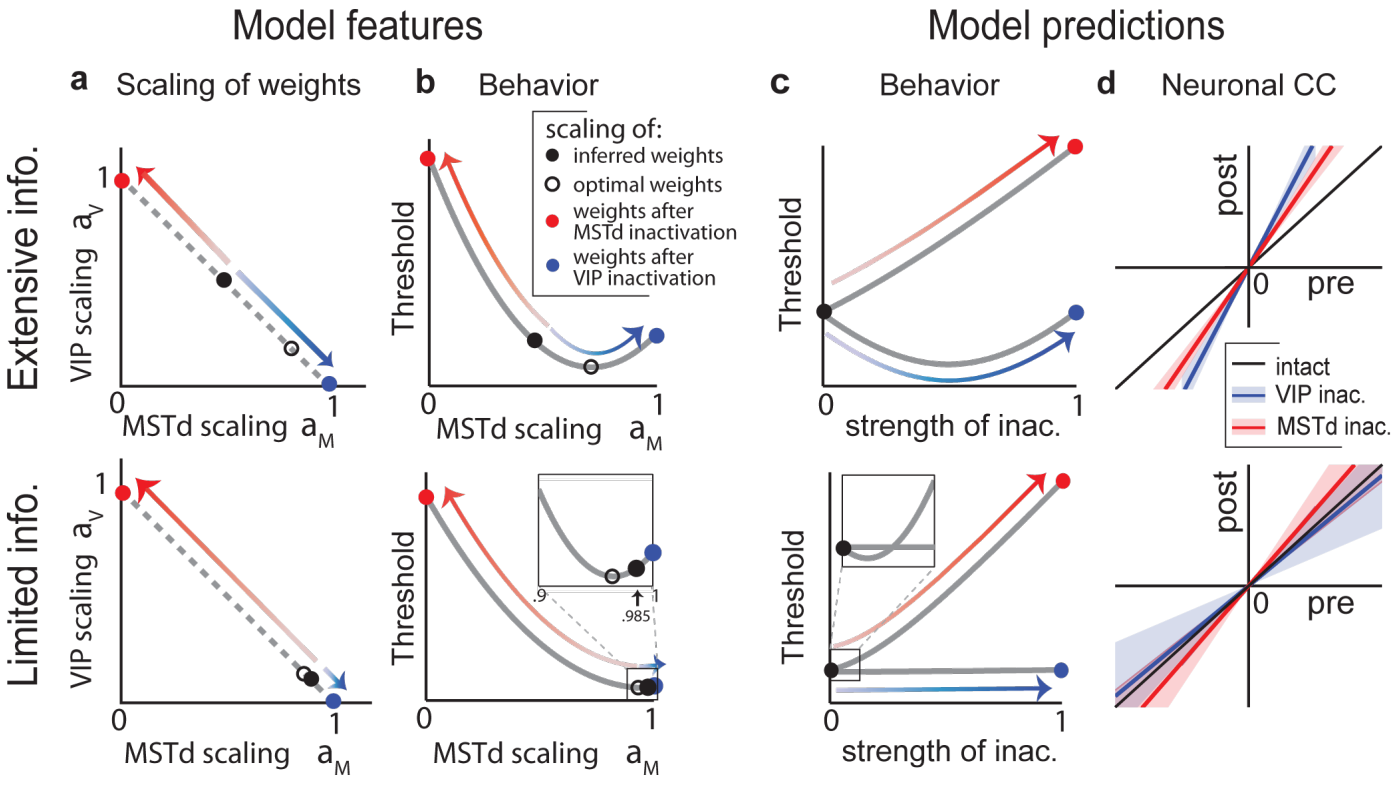

Figure 7. Decoding strategy and model predictions for the extensive information model and the limited information model. (a) Optimal (open black) and inferred (filled black) scaling of weights in MSTd (a_M) and VIP (a_V). Inactivation of either MSTd (red) or VIP (blue) confines the readout to the active 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 ϑ as a function of a_M. Whereas ϑ 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

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 ζ_M and ζ_V in **Equation (9)**, and shaded regions indicate ±1 s.d. of uncertainty.

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 **Figure 1d** and **Figure S4b** 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 (**Figure 7c – 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 (**Figure 7c – 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 (**Figure 7c**, 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.

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The decoding strategies implied by the two models also have different consequences for how CCs should change during inactivation experiments (**Methods M10**). 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 (**Figure 7d – top**). In the limited information model, inactivating either area produces no significant changes in the other's CCs (**Figure 7d – 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

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Equation 3.1 and 3.2**) 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 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 decision, 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

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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[38]. However, in simple tasks involving binary choices, and in areas in

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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[29,39–41]. 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[38], 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 (**Figure 7**). Future experiments will test each

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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[42]. In that work, the authors used a recurrent network model with mutual inhibition between populations[43,44] 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[45], and psychophysical kernels are sometimes skewed towards one end of the trial[46,47], suggesting that decoding could also be suboptimal in time. 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[46,48,49]. 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[46,48,49]. 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

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 (**Equation 2.1**). 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 (**Figure 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 (**Figure S15**). 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[1,46].

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 (**Figure 5c, 6a**) than is observed experimentally (**Figure 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[50]. 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.

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

METHODS

M1. Choice correlations in a linear feedforward model. Consider a standard feedforward decision process in which the neural response $\mathbf{r}\sim\mathcal{N}(\mathbf{f},\Sigma)$ is read out with weights $\mathbf{w}$ to generate an estimate $\hat{s}=\mathbf{w}^T(\mathbf{r}-\mathbf{f}(s_0))$. Choice correlations $\mathbf{C}$ in this scheme were previously shown[14,15] to be related to neuronal weights and response covariance according to $\mathbf{C}=S^{-1}\Sigma\mathbf{w}/\sqrt{\mathbf{w}^{\mathrm{T}}\Sigma\mathbf{w}}$ where $S=\sqrt{\mathrm{diag}(\Sigma)}$. We can decompose these choice correlations into a sum of components arising from the individual noise modes of the *N*x*N* covariance matrix Σ as: $\mathbf{C}=\sum_{i=1}^{N}\beta_i\mathbf{C}_{\mathrm{opt}}^i$ where $\mathbf{C}_{\mathrm{opt}}^i$ is the component of choice correlations generated from noise fluctuations along the i^{th} mode when decoding weights $\mathbf{w}$ are optimal (**Supplementary notes S1, S2**). $\mathbf{C}_{\mathrm{opt}}^i$ depends on the shape of the i^{th} noise mode $\mathbf{u}^i$, the amplitude of the signal $\mathbf{f}'$ (the derivative of the neurons' tuning curves), and the optimal threshold ϑ according to:

$$\mathbf{C}_{\mathrm{opt}}^i=\vartheta\left(\mathbf{f}'^T\mathbf{u}^i\right)S^{-1}\mathbf{u}^i \tag{4}$$

If decoding is optimal, then multipliers $\beta_i\equiv 1$ so the choice correlation $C_{k,\mathrm{opt}}$ of neuron k becomes $\sum_{i=1}^{N}C_{k,\mathrm{opt}}^i=\vartheta\sum_{i=1}^{N}\left(\mathbf{f}'^T\mathbf{u}^i\right)(S^{-1}\mathbf{u}^i)_k$ which reduces to ϑ/ϑ_k (**Supplementary note S2**) in agreement with earlier work[15]. In general however, multipliers β_i will be different from 1 and can be estimated by regressing measured choice correlations $\mathbf{C}$ against the corresponding component $\mathbf{C}_{\mathrm{opt}}^i$.

M2. Weight scaling factors for unbiased decoding. Let $\mathbf{w}=(a_x\mathbf{w}_x,a_y\mathbf{w}_y)^T$ denote the readout weights of neurons where a_x and a_y represent the scaling of weights in the two populations *x* and *y*. To ensure unbiased decoding both before and after inactivation of the individual populations *x* or *y*, $\mathbf{w}^T\mathbf{f}'$, $\mathbf{w}_x^T\mathbf{f}_x'$, and $\mathbf{w}_y^T\mathbf{f}_y'$ must all be equal to 1 where $\mathbf{f}'=(\mathbf{f}_x',\mathbf{f}_y')^T$ denotes the derivatives of the tuning curves of neurons in *x* and *y* (**Supplementary note S0**). This yields the constraint that $a_x+a_y=1$ at all times.

M3. Relation between behavioural threshold and weight scaling factors. Behavioural threshold ϑ is proportional to the square root of the decoder variance (with proportionality of 1 for threshold of 68% correct), so $\vartheta^2=\mathbf{w}^T\Sigma\mathbf{w}$. If decoding is confined to the subspace of leading eigenmodes of Σ spanned by neurons within *x* and *y* ($\mathbf{u}^x$ and $\mathbf{u}^y$), then $\mathbf{w}_x\propto a_x\mathbf{u}^x$ and $\mathbf{w}_y\propto a_y\mathbf{u}^y$ where the constants of proportionality are chosen to ensure unbiased decoding. 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 (**Supplementary note S4**):

$$\vartheta^2=a_x^2\varepsilon_{xx}+a_y^2\varepsilon_{yy}+2a_xa_y\varepsilon_{xy} \tag{5}$$

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where ε_{xx} and ε_{yy} are the magnitudes of noise within x and y, and ε_{xy} is the magnitude of correlated noise. Thresholds following inactivation can be determined by setting the weight scaling factor for the inactivated area to zero, yielding $\vartheta_{-x}^2 = \varepsilon_{yy}$ and $\vartheta_{-y}^2 = \varepsilon_{xx}$.

M4. 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 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 NIH guidelines. 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*, **Figure 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.

M5. 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].

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

M6. 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 μ and standard deviation ϑ, and this standard deviation defined the psychophysical threshold, corresponding to 68% correct performance (d'=1, assuming no bias, i.e. $\mu = 0^{\circ}$).

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:

$$\vartheta_k = 1/\sqrt{J_k} = \sigma_k / f_k' \tag{6}$$

where ϑ_k and J_k are the threshold and linear Fisher information[51] for neuron k, f_k' is the derivative of the neuron's tuning curve at the reference stimulus (0˚), and σ_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[52] (**Supplementary Figure 3**).

M7. 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[53]. 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[54]. 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 (**Supplementary note S0**). 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

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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] **(Equation 2.1)** 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 Supplementary notes 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).

M8. 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, $R_{ij}^{\mathrm{sig}} = \mathrm{Corr}(f_i, f_j)$, 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, $\overline{R}$, were then modeled as linearly proportional to the signal correlations:

$$\overline{R}_{ij} = (1 - m)\delta_{ij} + mR_{ij}^{\mathrm{sig}} \tag{7.1}$$

where δ_{ij} is the Kronecker delta function (δ_{ij} is 1 when *i*=*j*, and 0 otherwise) and *m* is the slope of the relationship between signal correlations and noise correlations. This slope was much steeper in VIP than MSTd[21]. For the vestibular condition, slopes were found to be m_M=0.19±0.08 and m_V=0.70±0.16 within MSTd and VIP respectively, and for the visual condition they were m_M=0.12±0.09 and m_V=0.50±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 $\overline{R}$ given by **equation 7.1** 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 Σ 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:

$$m_{MV} = k\sqrt{m_M m_V} \tag{7.2}$$

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

M9. 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:

$$\Sigma_{\mathrm{IL}} = \Sigma + \begin{bmatrix} \varepsilon_{MM}\mathbf{f}'_M\mathbf{f}'^{\,T}_M & \varepsilon_{MV}\mathbf{f}'_M\mathbf{f}'^{\,T}_V \\ \varepsilon_{MV}\mathbf{f}'_V\mathbf{f}'^{\,T}_M & \varepsilon_{VV}\mathbf{f}'_V\mathbf{f}'^{\,T}_V \end{bmatrix} \tag{8}$$

where $\mathbf{f}'_M$ and $\mathbf{f}'_V$ are derivatives of tuning curves of the neurons in M and V respectively, and Σ is the noise used in the extensive information model. Whereas $\mathbf{f}'_M$ and $\mathbf{f}'_V$ can be estimated by measuring the tuning curves of individual neurons, precisely estimating ε_{MM}, ε_{VV}, and ε_{MV} is difficult even with large-scale recordings as their magnitudes may be very small compared to the magnitude of noise in Σ. 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 **equation 5** where M and V take the places of x and y. Consequently, we can determine ε_{MM} and ε_{VV} from behavioural thresholds following inactivation using $\varepsilon_{MM} = \vartheta^2_{-V}$ and $\varepsilon_{VV} = \vartheta^2_{-M}$. We can then use **equation 5** in conjunction with **equation 3.2** to determine both the ratio a_M/a_V of weight scalings and the magnitude of correlation between populations $\varepsilon_{MV} = \gamma\varepsilon_{MM}$.

M10. 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 $\tilde{\mathbf{C}}_y$ 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 ζ_y and ζ_y are (**Supplementary note S10**):

$$\zeta_x = \frac{1}{\beta_x}\frac{\vartheta_{-y}}{\vartheta} \quad \text{and} \quad \zeta_y = \frac{1}{\beta_y}\frac{\vartheta_{-x}}{\vartheta} \tag{9}$$

where β_x and β_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.

Acknowledgements: The work was supported by NIH R01 DC04260, R21 DC014518 and the Simons Collaboration for the Global Brain, grant #324143. A.P. was supported by a grant from Simons Global Brain Initiative and the Swiss National Foundation (#31003A_143707). 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.

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Nienborg H, R. Cohen M, Cumming BG (2012) Decision-Related Activity in Sensory Neurons: Correlations Among Neurons and with Behavior. Annu Rev Neurosci 35: 463–483. doi:10.1146/annurev-neuro-062111-150403.

2. Georgopoulos AP, Schwartz AB, Kettner RE (1986) Neuronal population coding of movement direction. Science 233: 1416–1419. doi:10.1126/science.3749885.

3. Paradiso MA (1988) A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. Biol Cybern 58: 35–49.

4. Pouget A, Thorpe SJ (1991) Connectionist Models of Orientation Identification. Conn Sci 3: 127–142.

5. Seung HS, Sompolinsky H (1993) Simple models for reading neuronal population codes. Proc Natl Acad Sci U S A 90: 10749–10753. Available: http://www.ncbi.nlm.nih.gov/pubmed/8248166.

6. Shadlen MN, Britten KH, Newsome WT, Movshon JA (1996) A computational analysis of the relationship between neuronal and behavioral responses to visual motion. J Neurosci 16: 1486–1510. Available: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=8778300\nhttp://www.ncbi.nlm.nih.gov/pubmed/8778300.

7. Oram MW, Földiák P, Perrett DI, Sengpiel F (1998) The “Ideal Homunculus”: decoding neural population signals. Trends Neurosci 21: 259–265. doi:10.1016/S0166-2236(97)01216-2.

8. Chen Y, Geisler WS, Seidemann E (2006) Optimal decoding of correlated neural population responses in the primate visual cortex. Nat Neurosci 9: 1412–1420.

9. Cohen MR, Newsome WT (2009) Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. J Neurosci 29: 6635–6648.

10. Graf ABA, Kohn A, Jazayeri M, Movshon JA (2011) Decoding the activity of neuronal populations in macaque primary visual cortex. Nat Neurosci 14: 239–245.

11. Berens P, Ecker AS, Cotton RJ, Ma WJ, Bethge M, et al. (2012) A Fast and Simple Population Code for Orientation in Primate V1. J Neurosci 32: 10618–10626. doi:10.1523/JNEUROSCI.1335-12.2012.

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

12. Gu Y, Angelaki DE, DeAngelis GC (2014) Contribution of correlated noise and selective decoding to choice probability measurements in extrastriate visual cortex. Elife.

13. Crapse TB, Basso MA (2015) Insights into Decision-Making Using Choice Probability. J Neurophysiol: jn.00335.2015. Available: http://jn.physiology.org/lookup/doi/10.1152/jn.00335.2015.

14. Haefner RM, Gerwinn S, Macke JH, Bethge M (2013) Inferring decoding strategies from choice probabilities in the presence of correlated variability. Nat Neurosci 16: 235–242. Available: http://www.ncbi.nlm.nih.gov/pubmed/23313912. Accessed 18 September 2013.

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

16. Hanks TD, Kopec CD, Brunton BW, Duan CA, Erlich JC, et al. (2015) Distinct relationships of parietal and prefrontal cortices to evidence accumulation. Nature 520: 220–223. Available: http://www.ncbi.nlm.nih.gov/pubmed/25600270.

17. Raposo D, Kaufman MT, Churchland AK (2014) A category-free neural population supports evolving demands during decision-making. Nat Neurosci 17: 1784–1792.

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

19. Katz L, Yates J, Pillow JW, Huk AC (2016) Dissociated functional significance of decision-related activity in the primate dorsal stream. Nature 535: 285–288.

20. Moreno-Bote, R. B. J, Kanitscheider I, Pitkow X, Latham PE, Pouget A (2014) Information-limiting correlations. Nat Neurosci 17: 1410–1417.

21. Chen A, Deangelis GC, Angelaki DE (2013) Functional specializations of the ventral intraparietal area for multisensory heading discrimination. J Neurosci 33: 3567–3581. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3727431&tool=pmcentrez&rendertype=abstract. Accessed 22 December 2013.

22. Gu Y, DeAngelis GC, Angelaki DE (2012) Causal Links between Dorsal Medial Superior Temporal Area Neurons and Multisensory Heading Perception. J Neurosci 32: 2299–2313.

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

doi:10.1523/JNEUROSCI.5154-11.2012.

23. Gu Y, Angelaki DE, Deangelis GC (2008) Neural correlates of multisensory cue integration in macaque MSTd. Nat Neurosci 11: 1201–1210. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2713666&tool=pmcentrez&rendertype=abstract. Accessed 11 December 2013.

24. Advani M, Ganguli S (2016) Statistical Mechanics of Optimal Convex Inference in High Dimensions. Phys Rev X 6: 031034. Available: http://link.aps.org/doi/10.1103/PhysRevX.6.031034.

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

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

27. Sompolinsky H, Yoon H, Kang K, Shamir M (2001) Population coding in neuronal systems with correlated noise. Phys Rev E 64. doi:10.1103/PhysRevE.64.051904.

28. Averbeck BB, Lee D (2006) Effects of noise correlations on information encoding and decoding. J Neurophysiol 95: 3633–3644.

29. Gu Y, Liu S, Fetsch CR, Yang Y, Fok S, et al. (2011) Perceptual learning reduces interneuronal correlations in macaque visual cortex. Neuron 71: 750–761.

30. Liu S, Gu Y, DeAngelis GC, Angelaki DE (2013) Choice-related activity and correlated noise in subcortical vestibular neurons. Nat Neurosci 16: 89–97. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3612962&tool=pmcentrez&rendertype=abstract.

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

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

33. Shamir M, Sompolinsky H (2006) Implications of neuronal diversity on population coding. Neural

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Comput 18: 1951–1986. doi:10.1162/neco.2006.18.8.1951.

34. Ecker AS, Berens P, Tolias AS, Bethge M (2011) The Effect of Noise Correlations in Populations of Diversely Tuned Neurons. J Neurosci 31: 14272–14283. doi:10.1523/JNEUROSCI.2539-11.2011.

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

36. Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding and computation. Nat Rev Neurosci 7: 358–366.

37. Schneidman E, Bialek W, II MJB (2003) Synergy, Redundancy, and Independence in Population Codes. J Neurosci 23: 11539–11553.

38. Beck JM, Ma WJ, Pitkow X, Latham PE, Pouget A (2012) Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. Neuron 74: 30–39.

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

40. Jehee JFM, Ling S, Swisher JD, van Bergen RS, Tong F (2012) Perceptual learning selectively refines orientation representations in early visual cortex. J Neurosci 32: 16747–16753a. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3575550&tool=pmcentrez&rendertype=abstract.

41. Li W, Piëch V, Gilbert CD (2004) Perceptual learning and top-down influences in primary visual cortex. Nat Neurosci 7: 651–657.

42. Kopec CD, Erlich JC, Brunton BW, Deisseroth K, Brody CD (2015) Cortical and Subcortical Contributions to Short-Term Memory for Orienting Movements. Neuron 88: 367–377.

43. Wong K-F, Wang X-J (2006) A recurrent network mechanism of time integration in perceptual decisions. J Neurosci 26: 1314–1328. Available: http://www.jneurosci.org/content/26/4/1314.full.

44. Machens CK, Romo R, Brody CD (2005) Flexible Control of Mutual Inhibition: A Neural Model of Two-Interval Discrimination. Science (80-) 307: 1121–1124. Available: http://www.sciencemag.org/cgi/content/abstract/307/5712/1121\nhttp://www.ncbi.nlm.nih.gov/pubmed/15718474.

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

45. Park IM, Meister MLR, Huk AC, Pillow JW (2014) Encoding and decoding in parietal cortex during sensorimotor decision-making. Nat Neurosci 17: 1395–1403. Available: http://dx.doi.org/10.1038/nn.3800.

46. Nienborg H, Cumming BG (2009) Decision-related activity in sensory neurons reflects more than a neuron's causal effect. Nature 459: 89–92.

47. de Lafuente V, Jazayeri M, Shadlen MN (2015) Representation of accumulating evidence for a decision in two parietal areas. J Neurosci 35: 4306–4318. Available: http://www.jneurosci.org/content/35/10/4306.full.

48. Yang H, Kwon SE, Severson KS, O'Connor DH (2015) Origins of choice-related activity in mouse somatosensory cortex. Nat Neurosci 19: 127–134. Available: http://www.nature.com/neuro/journal/v19/n1/full/nn.4183.html?WT.ec_id=NEURO-201601&spMailingID=50354006&spUserID=Njk2Njk2MzE4MjUS1&spJobID=824237578&spReportId=ODI0MjM3NTc4S0.

49. Wimmer K, Compte A, Roxin A, Peixoto D, Renart A, et al. (2015) Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. Nat Commun 6: 6177. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4347303&tool=pmcentrez&rendertype=abstract.

50. Gao P, Ganguli S (2015) On simplicity and complexity in the brave new world of large-scale neuroscience. Curr Opin Neurobiol 32: 148–155.

51. Beck J, Pouget A (2011) Insights from a Simple Expression for Linear Fisher Information in a Recurrently Connected Population of Spiking Neurons. Neural Comput 23: 1484–1502.

52. Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. J Neurosci 12: 4745–4765. doi:10.1.1.123.9899.

53. Green DM, Swets JA (1966) Signal detection theory and psychophysics. Wiley. p. 174 p. Available: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Signal+detection+theory+and+psyc

bioRxiv preprint doi: https://doi.org/10.1101/108019; this version posted May 22, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

hophysics#0.

54. Kang I, Maunsell JHR (2012) Potential confounds in estimating trial-to-trial correlations between neuronal response and behavior using choice probabilities. J Neurophysiol.