Navigating the Neural Space in Search of the Neural Code

Mehrdad Jazayeri[1,*] and Arash Afraz[1]
[1]Department of Brain & Cognitive Sciences, McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*Correspondence: mjaz@mit.edu
http://dx.doi.org/10.1016/j.neuron.2017.02.019

The advent of powerful perturbation tools, such as optogenetics, has created new frontiers for probing causal dependencies in neural and behavioral states. These approaches have significantly enhanced the ability to characterize the contribution of different cells and circuits to neural function in health and disease. They have shifted the emphasis of research toward causal interrogations and increased the demand for more precise and powerful tools to control and manipulate neural activity. Here, we clarify the conditions under which measurements and perturbations support causal inferences. We note that the brain functions at multiple scales and that causal dependencies may be best inferred with perturbation tools that interface with the system at the appropriate scale. Finally, we develop a geometric framework to facilitate the interpretation of causal experiments when brain perturbations do or do not respect the intrinsic patterns of brain activity. We describe the challenges and opportunities of applying perturbations in the presence of dynamics, and we close with a general perspective on navigating the activity space of neurons in the search for neural codes.

Neuroscience research is undergoing a transformation of scale along two axes. On the one hand, experiments in the age of "-omics" (genomics, proteomics, connectomics, etc.) are providing information on large-scale correlations at multiple scales, from genes to behavior. On the other hand, increasingly more precise perturbation techniques are beginning to reveal causal relationships at an unprecedented level of detail. The time is ripe for a fresh perspective on how to optimally harness these data toward a greater understanding of the brain and behavior. Here we evaluate the strengths and weaknesses of measurement and perturbation experiments in terms of their ability to generate and refine theories and causal models of brain function.

Correlation and Causation: A Brief Primer

Correlational dependencies describe associations that we measure but do not control (Figure 1A), whereas causal dependencies link a dependent variable to an experimentally controlled variable (Figures 1A and 1B). Note that we use the term correlation to refer to any form of statistical dependence. In neuroscience research, experiments that rely on measurements of neural activity (e.g., extracellular recordings) are commonly dubbed correlational, and experiments that involve direct perturbation of neural activity (e.g., microstimulation), causal. Intuitively appealing as it may be, this connotation is misleading. The key concept in causal inference is randomization, i.e., setting the value of a variable independent of other variables. Experiments that only rely on measurements of brain activity may indeed reveal causal dependencies if they employ a randomized variable, such as an external stimulus. For example, we can confidently state that a neural response is caused by a sensory stimulus, if the sensory input is randomized. This logic, however, can belie the complexity of the causal chain when the causal chain between the stimulus and the neural response is long. For example, a contrived experiment might establish a causal link between neural activity in motor neurons and a looming threat. However, numerous intervening nodes between the cause and effect would limit the generality of the inferences one can make from such causal observation about the function of motor neurons.

Conversely, inferences one can make from direct perturbations of brain activity are not always causal. Dependencies that involve a randomized variable are causal, whereas dependencies among all non-randomized variables are correlational. Although this seems like a straightforward distinction, in practice it may be difficult to ascertain which variables are directly randomized. We will use two hypothetical examples to crystallize the nuanced relationship between types of inference (causal versus correlational) and experimental techniques (perturbation versus measurement).

Example 1

An interventional study uses marmoset to examine the neural basis of auditory spatial perception. The experimenter records neural activity along the auditory pathway while an animal discriminates the azimuth of a randomly positioned sound source. In this experiment, one can assess four types of relationships: (1) the relationship between stimulus and behavior, (2) the relationship between stimulus and neural activity, (3) the relationship between neural activity and behavior, and (4) the relationship between neural activity in two brain areas. The inferences one can make in the first two cases are causal because one of the variables (i.e., the stimulus) was randomized. In contrast, the last two relationships are correlational because neither the neural activity nor the behavior was independently randomized. This

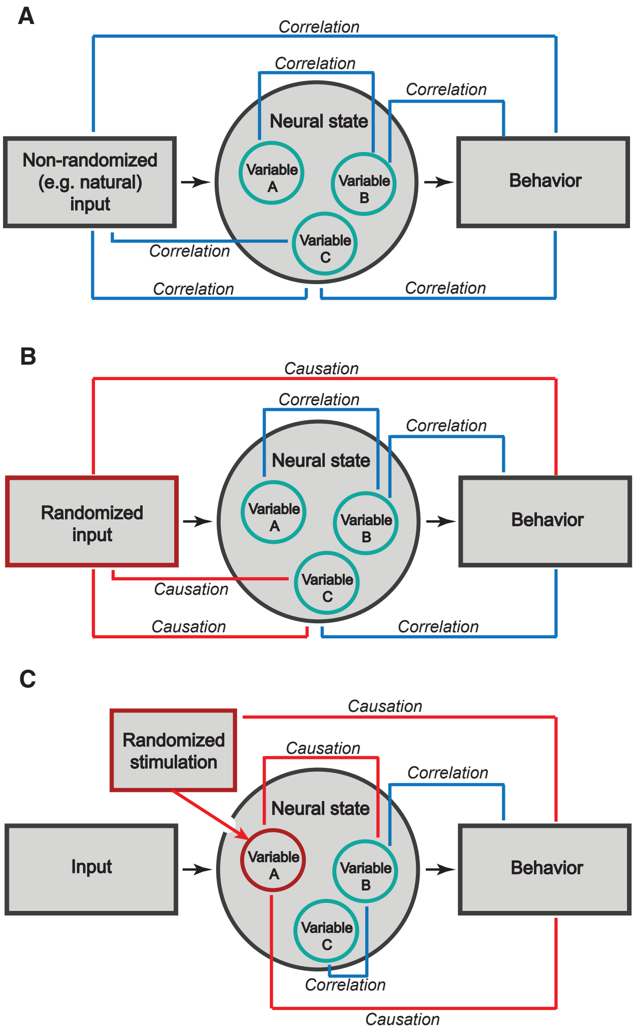

Figure 1. Randomization, Correlation, and Causation Using Measurements and Perturbations

(A) A completely observational study that involves recording neural activity and behavior of a free-range animal in its natural habitat. Any statistical dependency observed under such circumstances is correlational because no variable is externally randomized. Here even sensory responses to natural stimuli do not reflect causation because the appearance of the stimulus is determined by the animal and other factors that are not controlled or randomized by the experimenter.

(B) An experiment in which the input (e.g., auditory stimulus) is experimentally randomized (e.g., sound played at random times), and the ensuing brain activity is measured. Here any aspect of brain activity (e.g., activity in the auditory cortex) and/or behavior (e.g., orienting) that depends on the stimulus reveals a causal relationship. However, relationships between the activity of different variables in the brain and the dependency of the behavior on the brain remain correlational.

(C) An experiment employing a perturbation technique (e.g., microstimulation). Here again the choice of correlational versus causal inference depends not on the experimental technique but on the variables of interest. The relationship between every dependent variable and the randomized variable is causal. The relationship between all variables that are not randomized (e.g., activity in the brain areas that are not directly perturbed), as well as the relationship between non-randomized variables and behavior, remains correlational.

experiment can be used to make causal inferences about the nature of stimulus encoding by auditory neurons as well as the effect of the stimulus on behavior. However, statements about whether and what features of the neural responses support the behavior remain correlational.

Example 2

An interventional study in a mouse model aims to test the causal role of a projection from area A to area B in a specific behavior. Using advanced optogenetic techniques, the experimenter selectively randomizes the spiking activity of area A neurons that project to area B, and the experimenter makes the two unambiguous causal inferences that the firing rate of neurons in area A that project to area B can (1) control firing rates in area B and (2) influence behavior. However, the link between area B activity and behavior remains correlational, unless the experimenter can rule out other secondary and/or off-target effects downstream of the optical stimulation.

As a final point, we note that the empirical assessment of causal statements is not as straightforward as their theoretical definition. Causal statements imply that we can directly control the variables of interest. However, many quantities we may wish to perturb, such as the firing rate of a neuron, are *latent variables* that are not accessible directly. For latent variables as well as other variables that are technically difficult to perturb, causal statements must be viewed in shades of gray, depending on the extent to which we can control the variable deterministically. For example, in Figure 1C, if variable A is not deterministically controlled by the randomized stimulation, we may not be able to ascertain that the dependency of variable B on variable A is causal.

Correlation and Causation in a Multi-scale System

Correlational statements do not imply causation, but causal statements are not always revealing either. As Weiskrantz wrote in *Analysis of Behavioral Change*, “a simple statement that Task X is affected by Treatment A is inadequate except insofar as a confession that one has a research program” (Weiskrantz, 1968). The logic of this rather harsh characterization is that behaviors typically rely on numerous entangled capacities that are difficult to tease apart. An old riddle might clarify this point. If removing a transistor from a radio adds noise to the sound, can we conclude that the transistor’s function is to improve signal-to-noise? Similarly, if silencing a specific cell type increases reaction time in a certain behavior, can we conclude that it controls reaction time? The answer to both questions may be positive, but such perturbations might also lead to incorrect inferences.

Nearly two decades ago, in a thought-provoking paper entitled “Can Molecules Explain Long-Term Potentiation?”, Lichtman and Sanes argued that long-term potentiation (LTP) may not be straightforwardly explained in terms of its underlying molecular causes (Sanes and Lichtman, 1999). They noted several challenges: (1) many molecular causes are better construed as modulators of LTP and do not necessarily mediate it, (2) off-target pathways can lead to spurious causal effects, (3) perturbations can interact with uncontrolled variables in unpredictable ways, and (4) the brain is extremely complex and not all cellular phenomena find a coherent explanation in terms of the full list of the underlying molecular components. The first three points are essential for any well-designed causal experiment. The last point identifies a deeper challenge: explaining a phenomenon expressed at the cellular level in terms of interactions at the molecular level.

This tension between levels of analysis looms large in neuroscience, and is evident across multiple spatial and temporal scales, from molecules to neurons, from neuron to neural circuits, and from neural circuits to behavior. The challenge arises from attempting to assess the causes of a phenomenon observed at one scale using perturbation tools that operate at another scale. One way such a level mismatch could occur is when the phenomenon of interest operates below the resolution of our perturbation tool. For example, appropriate techniques do not yet exist to selectively perturb distinct mechanisms of dendritic integration in vivo. In these scenarios, the most direct solution might be to develop tools that can target the system at progressively higher levels of spatial and temporal resolution with sufficient specificity and reliability. This is a great challenge but the solution is likely to come in time, as hinted by the momentum of technological advances in targeting cell types, genes, and proteins (Adamala et al., 2016; Cong et al., 2013; Gradinaru et al., 2010; Klapoetke et al., 2014; Marshel et al., 2010).

Another important form of mismatch occurs when perturbations are able to push and pull the low-level components of the system every which way, but the behavior depends on the collective and coordinated interaction of those components. In theory, this form of mismatch does not pose a fundamental limitation. After all, if we can randomize the variables associated with individual components, we may be able to randomize their collective behavior as well by applying combinations of perturbations (Klapoetke et al., 2014; Prigge et al., 2012). This strategy might require searching in a large space but this is not necessarily a problem. For example, in a recent study in larval *Drosophila*, high-throughput random perturbations were used to map specific movement motifs onto specific subsets of neurons (Vogelstein et al., 2014). However, this bottom-up strategy needs to address two important requirements: (1) reasonably low-cost and high-throughput screening tools, and (2) statistical and machine learning tools that can handle the dauntingly large datasets that such an approach generates.

The same approach may be less effective when the search space becomes prohibitively large, as is the case when dealing with more complex nervous systems and more integrative and flexible behaviors. For example, we would be hard pressed to make a case for a one-to-one causal map between specific genes or a small number of neurons and the ability to perceive, move, or perform cognitive tasks. Such high-level functions typically rely on structured interactions between large groups of neurons and neural circuits (Buschman and Miller, 2007; Churchland et al., 2012; DiCarlo and Cox, 2007; Georgopoulos, 1994; Hanks et al., 2015; Harvey et al., 2012; Jazayeri and Movshon, 2006; Laurent et al., 2001; Lo and Wang, 2006; Mante et al., 2013; Ma et al., 2006; Moore and Armstrong, 2003; Raposo et al., 2014; Rigotti et al., 2013; Siegel et al., 2015; Znamenskiy and Zador, 2013), and they are thus difficult to explore by searching through the combinatorial space of possible low-level perturbations.

Cogwheels in a Clock and Neural Codes in the Brain

Imagine a Martian aiming to study the inner workings of a human-made clock using perturbation tools that can move around every atom inside the clock. Although the clock is made of atoms, it would take a gargantuan effort to understand the logic of the most basic mechanical interactions inside the clock using analyses at the atomic level. Perhaps a more suitable approach would be to perturb the clock at the level of its key functional components, such as its cogwheels and springs. To perturb a cogwheel as a whole, the Martian has to take two constraints into account: (1) respect the integrity of the cogwheel by not moving the atoms of the cogwheel relative to one another, and (2) randomize the position of the cogwheel by moving the atoms of the cogwheel together while respecting their relative positions.

To understand how these constraints apply to causal experiments in neuroscience, let us consider individual neurons as the atoms of the system. The main ideas in our discussion are scalable and do not depend on the level at which the atoms are defined, but considering the neuron as the atom will help in developing the intuition. Imagine a subset of neurons whose activity causally and selectively drives a behavior of interest. This is similar to a perturbation in the clock that is able to take control of the movements of clock hands, and it would mean that this subset contains the key features (i.e., cogwheels) that drive the behavior. We will refer to the minimum set of features that control the behavior as the neural code (Johnson, 2000).

How can we use perturbations to discover the neural code? Random perturbations of neurons are analogous to randomly moving every atom of the cogwheel, and they could limit our ability to see the forest for the trees. Just like the Martian, a more efficient path might be for the perturbations to respect the structure of neural activity among those neurons that contain the neural code. Here we propose a conceptual framework based on geometric representations of neural activity and behavior that may facilitate the process of making inferences about the results of perturbation experiments in terms of the neural code.

Intrinsic Neural Manifold

What constrains a cogwheel is the relative position of its atoms. In the brain, this constraint applies to the structure of responses across subgroups of neurons that control certain desired aspects of a behavior. We refer to the activity patterns of neurons during a desired behavioral task and in response to a well-defined stimulus set as the intrinsic patterns of activity. Since neural responses are constrained by anatomy and by their inputs, intrinsic activity is usually a small subset of all the possible activity patterns. We can visualize the intrinsic patterns of activity within a coordinate system in which each axis corresponds to the activity of one neuron. We refer to the space spanned by the axes as the state space and the parts of the space that correspond to the observed activity patterns as the intrinsic manifold (Figure 2). We use the term manifold in its general sense and do not imply any specific constraint on its continuity or topology. For example, a three-neuron network (Figure 2B) would correspond to a three-dimensional state space (Figure 2A). This representation divides the space of all possible activity patterns to two subspaces, one that is associated with the native activity patterns of the neurons under study (points on the manifold) and one that those neurons do not visit (the rest of the state space).

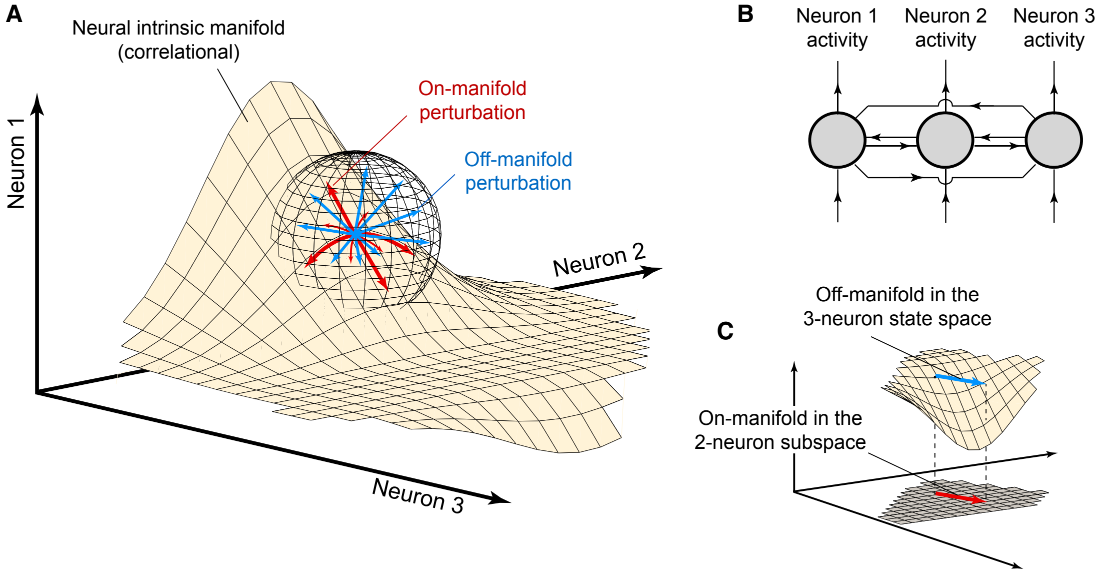

Figure 2. Intrinsic Neural Response Manifold and Perturbation Experiments

(A) The coordinate system represents the space of all possible activity patterns across *N* neurons (shown for three neurons). Not all combinations of neural response patterns occur during a certain behavioral task. The surface represents the subspace of activity patterns of those *N* neurons measured during a behavioral task, which we refer to as the intrinsic manifold. Perturbations are shown as arrows emanating from a point on the intrinsic manifold. Perturbations that respect the correlational structure of the neural activity land on the response manifold, and they are referred to as on-manifold perturbations (red arrows). Perturbations that push the system away from the intrinsic manifold are referred to as off-manifold perturbations (blue arrows).

(B) A schematic showing a network consisting of three neurons with all to all interactions, which would create an intrinsic manifold in a three-dimensional state space, as shown in (A). The shape of the manifold depends on the inputs to the three neurons as well as the constraints imposed by their interactions. The axes of the intrinsic manifold correspond to the activity (i.e., output) of the three neurons.

(C) The three-dimensional coordinate system and the colored surface show the state space and the intrinsic manifold for three neurons, respectively. The gray region shows the intrinsic manifold for neurons 2 and 3, which is a region within the two-dimensional state space spanned by those neurons. A perturbation on the gray surface would constitute an on-manifold perturbation for neurons 2 and 3. However, as shown by the blue arrow, the same perturbation could be off the higher-dimensional manifold that includes neuron 1. In other words, a perturbation that sets the state of the system on the intrinsic manifold of a subset of neurons (red arrow on the gray surface) could be off the intrinsic manifold with respect to other neurons (blue arrow off the colored surface). Stated in terms of neural activity patterns, if one perturbs without controlling for the activity of the third neuron, the overall perturbation might not respect the correlation structure of all three neurons.

Perturbing a clock requires that we respect the integrity of its cogwheels. The same can be said about the brain. The geometric representation of intrinsic manifolds provides a straightforward recipe for how this can be done: perturbations have to remain on the intrinsic manifold characterized by correlational measurements. The suggestion that a perturbation should remain on the intrinsic manifold may seem inconsistent with the idea of randomization, but it is not. Just like turning a specific cogwheel inside a clock, a perturbation can remain on manifold for a subset of neurons (i.e., at lower dimension) and still be off manifold with respect to the rest of the system (i.e., higher dimensions). Indeed, on-manifold perturbations that are restricted to a subset of neurons would automatically break correlations with neurons not included in the subset (Figure 2B). We will refer to the manifold associated with the subset of neurons under investigation as the neural manifold of interest (NMOI), and we emphasize that both on- and off-manifold perturbations with respect to NMOI are off manifold when viewed in a higher-dimensional state space that includes other neurons in the system.

Intrinsic Behavioral Manifold

To evaluate a causal link between neural activity and behavior, in addition to perturbing neural activity, we need to have a solid framework to analyze the corresponding behavioral outcomes. This is a critical aspect of causal experiments as many measurable behavioral outcomes of perturbations may be uninterpretable. For example, if perturbing a certain region of the basal ganglia were to alter movement kinematics, it would not necessarily follow that the basal ganglia is directly involved in the control of kinematics. Therefore, causal experiments are incomplete unless we have clear hypotheses about the space of possible behavioral outcomes. To characterize behavior, one has to choose a specific set of variables (position, speed, choice, reaction time, etc.), and one has to measure them during the experimental conditions of interest (i.e., in the absence of any perturbation). We can represent a set of desired behavioral variables during an experiment by an intrinsic behavioral manifold whose axes correspond to the measured variables (Figure 3). The form and dimensionality of this manifold depend on the experimental paradigm. As a general rule, the more complex the behavior is and the more aspects of behavior one wishes to explain, the higher dimensional the behavioral manifold would become. Similar to NMOI, we define the behavioral manifold of interest (BMOI) as the manifold that characterizes the behavioral variables that are under investigation. Naturally, one's success in examining the neural basis of behavior depends on a prudent choice of behavioral readout.

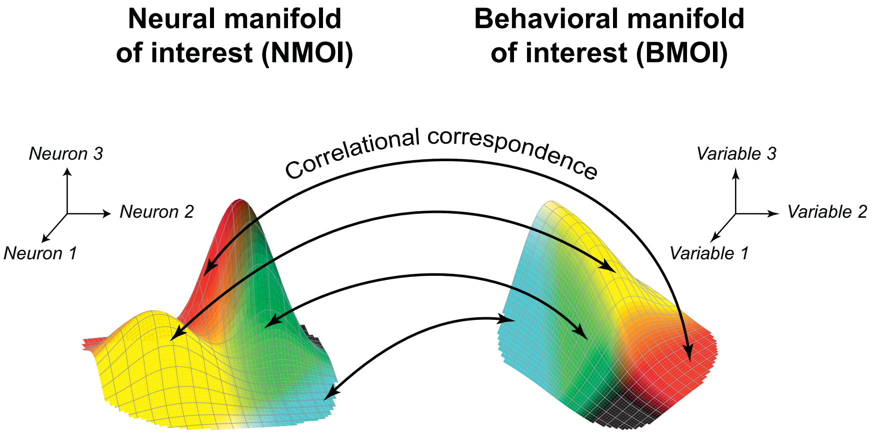
Neural manifold of interest (NMOI)
Behavioral manifold of interest (BMOI)
Correlational correspondence
Neuron 3
Neuron 2
Neuron 1
Variable 3
Variable 2
Variable 1

Figure 3. Correlational Correspondence between Intrinsic Neural Manifold and the Manifold of the Measured Behavioral Variables

The figure represents a hypothetical correlational experiment in which the experiment measures the activity of *N* neurons (shown for three neurons) and the corresponding value of *M* behavioral variables (shown for three variables). The left and right drawings represent the intrinsic neural and behavioral manifolds as low-dimensional surfaces within their respective 3D state spaces. The color code is arbitrary and is used in conjunction with the double-headed arrows to convey the correlational correspondence between subregions of the two manifolds that have the same color.

Knowledge about NMOI and BMOI and the correspondence between them provides a basis for assessing the behavioral outcome of perturbations. In what follows, we provide an in-depth analysis of what inferences one can make when either the neural perturbations or their behavioral outcomes remain on or move off the corresponding intrinsic manifolds.

On-Manifold Neural Perturbations

The simplest causal experiments to analyze are those in which the neural perturbations are on manifold (i.e., the perturbed state corresponds to a point on the NMOI). Again, recall that this is an on-manifold perturbation with respect to the NMOI and an off-manifold one with respect to the rest of the brain. An on-manifold perturbation can lead to one of several possible behavioral outcomes. One possibility is that the behavior associated with the perturbed state follows directly from the correlational link between the NMOI and BMOI (Figure 4A, a). This would indicate that the NMOI contains a neural code. However, it is possible that some of the neurons included in the NMOI are not necessary to drive the behavior. As such, selective control of behavior by an on-manifold perturbation invites a search for smaller subsets of neurons that might contain the key relevant features.

For example, let us assume that we successfully control a monkey's behavior in a face gender discrimination task by controlling the activity across the whole ventral visual pathway. We would then use on-manifold perturbations to target local regions within that pathway, such as the inferotemporal cortex, and, with success, we could move further to find neural clusters or cell types that continue to contain the neural code. The search for lower-dimensional neural manifolds continues until it is no longer possible to control the behavior selectively. At that stage, we may be close to the smallest dimensional neural manifold that contains the neural code. We note that the dimensionality of the neural code may still be smaller than the smallest NMOI. We will come back to this point in later sections.

Another possibility is that an on-manifold neural perturbation leads to no change in behavior (Figure 4A, c). This is, in effect, a negative result. Yet, because the perturbation was respectful of the intrinsic manifold, the negative result is meaningful. It suggests that either the correlation to behavior is epiphenomenal (i.e., does not contain the neural code) or perturbation of other or more neurons is needed to take control of the behavior. Therefore, in contrast to the previous case, an on-manifold perturbation with no behavioral effect invites a search for the neural code at higher dimensions.

On-manifold perturbations can also lead to other behavioral effects. For example, they may create non-selective changes in the behavior (Figure 4A, b) or push the behavior off the BMOI (Figure 4A, d). Although an exhaustive discussion of all possible behavioral outcomes is beyond the scope of this paper, we note that analyses along these lines could greatly enrich the interpretation of causal experiments.

Challenges in Performing On-Manifold Perturbations

In reality, our perturbation techniques are relatively crude. Not only do we not have precise control over the perturbation vector but also we rarely can fully characterize a perturbed state. Moreover, even when we have a reasonably precise perturbation tool, the system's internal constraints might redirect an intended on-manifold perturbation vector off the manifold. Let us use an example to clarify this point. Imagine an NMOI that corresponds to the coordinated activity of ten different neuron types in a brain area. Let us assume that we want to evaluate the causal effect of type 1 neuron on the behavior. To do so, we use a cell type-specific perturbation tool to randomly increase or decrease the activity of type 1 neurons. In this case, our intended perturbation vector is one that moves the state within the subspace spanned by neurons of type 1. What would be the direction of the resulting perturbation vector? Would it remain within the desired subspace or would the network interactions cause other neuron types to also change their activity? The answer depends on how the perturbation interacts with the system's intrinsic constraints. Imagine that perturbing type 1 neurons leads to a measurable behavioral effect. With this observation, we can safely argue that type 1 neurons causally influence the behavior. However, this statement can be difficult to interpret if the behavior is controlled by type 2 neurons that are influenced by perturbing type 1 neurons (Figure 5).

In general, two classes of factors can redirect an intended perturbation vector: off-target and secondary effects. Off-target effects correspond to changes of variables that the perturbation

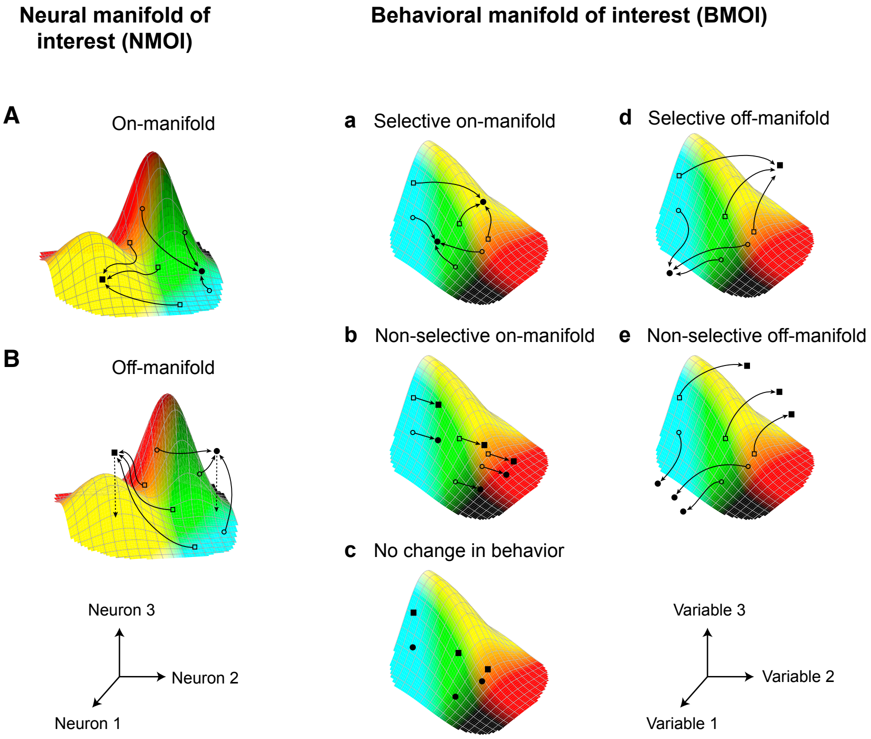

Figure 4. Behavioral Outcomes for On-Manifold and Off-Manifold Perturbations

(A and B) An intrinsic neural manifold of interest (NMOI) as a surface in a 3D coordinate system for three neurons. (A) This manifold with two sets of on-manifold perturbations (square and circle) is shown. For each set, the perturbations are shown as arrows moving the state of the system from arbitrary positions (open symbols) to a desired target state (filled symbols). (B) The same for two off-manifold perturbations is shown. The dotted arrows in (B) correspond to the projection of the off-manifold perturbations on the surface to convey the idea that these perturbations might lead to meaningful movements on the manifold. (a–e) The intrinsic behavioral manifold of interest (BMOI) as a surface in a 3D coordinate system for three measured behavioral variables. The colors, arrows, and symbols in each panel of the BMOI represent one type of behavioral outcome. (a) represents an outcome in which the two sets of perturbations can selectively control the behavioral variables on their intrinsic manifold. (b) represents an outcome in which the three converging neural perturbations do not lead to converging behavioral outcomes but the behavioral effects remain on the manifold. This would amount to a non-selective change in the behavior. As an example, (b) shows the case where all the perturbations cause a shift along a specific direction on the behavioral manifold. In practice, this could be a non-selective bias in behavior, or a non-selective increase in reaction time, or any other non-selective change in behavior. (c) represents the case where the neural perturbations are ineffective and have no measurable behavioral outcomes. (d) and (e) represent two cases where the perturbation pushes the behavior off its intrinsic manifold, by either causing an unexpected but selective behavioral outcome (d) or by causing a non-selective behavioral outcome (e). Both on- (A) and off-manifold (B) neural perturbations could lead to any of the hypothetical behavioral outcomes (a–e), and each combination of neural perturbation and behavioral outcome will have its own interpretation.

directly interacts with but are not the intended target (e.g., stimulating the axons of passage when microstimulating a cluster of neurons). Secondary effects are effects downstream of the intended target that may or may not be the key variables influencing the behavior (Otchy et al., 2015). Identifying and rectifying these unintended effects may be easy in simple systems with a few interacting nodes, but they would become challenging when the NMOI and BMOI are high dimensional.

One way to evaluate the potential concerns about off-target and secondary effects is to use causal belief networks (Figure 6). A belief network is a graphical model with nodes representing the variables of interest and arrows indicating presumed causal relationships between those variables. In this representation, the effect of randomization is unambiguous. Randomization dissociates the randomized variable (e.g., firing rate of cell type A) from all its *parents* (variables that causally influence it), but it allows it to exert its effect on all its *children* (variables it influences). This can be directly linked to our discussion of on- and off-manifold perturbations. When a variable is randomized, its parent nodes cannot exert any influence on the perturbation vector. However, the perturbation vector can be redirected along the axes that correspond to the children nodes of the randomized variable (second-order effect). In other words, the causal belief network and the intrinsic manifold provide a rigorous graphical and geometrical interpretation of how perturbations change the state of the system and how that change can be evaluated with respect to the native activity patterns in the system.

Another reason on-manifold perturbations may be challenging is that many behaviors we study in the laboratory are relatively low dimensional (i.e., BMOI is low dimensional). These behaviors are likely to rely on similarly low-dimensional patterns of neural activity (Brody et al., 2003; Churchland et al., 2012; Cunningham and Yu, 2014; Fitzgerald et al., 2013; Ganguli et al., 2008; Holdefer and Miller, 2002; Kaufman et al., 2014; Kobak et al., 2016; Li et al., 2016; Mante et al., 2013; Sadtler et al., 2014). Since our current perturbation tools cannot be tuned for a specific NMOI, when the manifold is low dimensional, random perturbations are more likely to target states that are off the manifold. Stated more quantitatively, the magnitude of the projection of a random N-dimensional vector onto its lower-dimensional subspaces drops rapidly as N gets larger. This could render a search via random perturbations ineffective even when the underlying neural manifold contains the neural code. This problem may be rectified by adopting more sensitive behavioral metrics or expanding the range of behavioral tasks and variables we monitor. This would increase the dimensionality of both the behavioral and neural manifolds. Perhaps the best example of this strategy comes from the groundbreaking cortical mapping experiments of Penfield and Jasper, where they directly asked subjects to report the percepts evoked by cortical stimulation (Penfield and Jasper, 1954). Penfield's experiments additionally benefited from the fact that the cortex has a modular organization. In other words, not only the behavioral readout was associated with a large intrinsic manifold but also the perturbation respected the

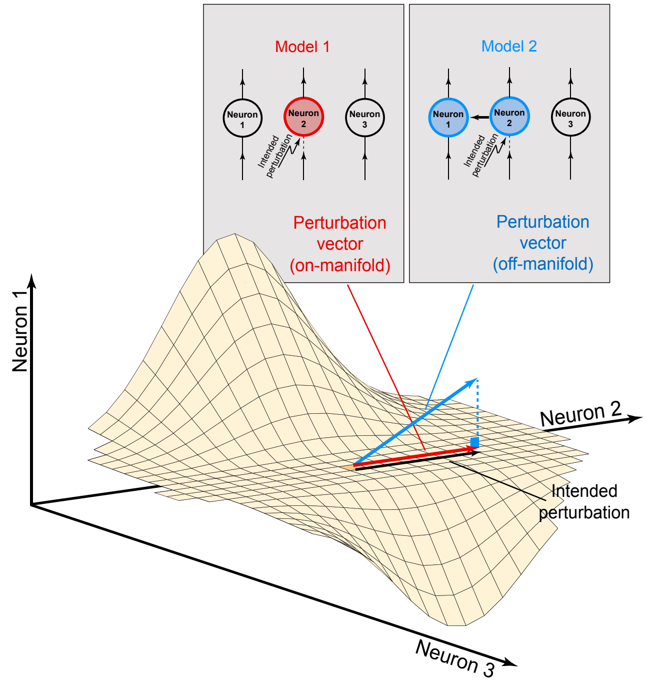

Figure 5. Second-Order Effects as Redirecting the Perturbation Vector

Model 1 represents a system of three independent neurons. The shape of the intrinsic manifold of this system is directly dictated by its input pattern. Stimulation of neuron 2 (black curved arrow) is intended to clamp the activity of neuron 2 to a high value independent of its input (black arrow on the manifold). Since the three neurons in model 1 do not interact, the actual perturbation vector (red arrow) would follow the intended direction (along the neuron 2 axis). Model 2 depicts a system with a lateral connection: neuron 2 activates neuron 1. The shape of the intrinsic manifold of this system is determined by a combination of its input pattern and the lateral interaction. The flat part of the manifold along the neuron 2 axis corresponds to a region where the activity of neuron 1 does not change (i.e., the two inputs to neuron 1 cancel out). In model 2, the same intended perturbation disrupts the balance of inputs to neuron 1 and leads to a perturbation vector (blue arrow) that has an unintended projection along the neuron 1 axis. The colored circles show the neurons that are affected by the perturbation in each of the two models.

coarse grain structure of the manifold and only went off it with respect to details that did not interfere with the behavioral outcome.

Behavioral Outcomes of Off-Manifold Perturbations

Let us first consider an off-manifold perturbation that alters behavior selectively (Figure 4B, a). Similar to the case of a selective on-manifold perturbation (Figure 4A, a), this finding suggests that the neural manifold contains a neural code, but it may provide more information than an on-manifold perturbation. It may additionally reveal that the dimensions along which the perturbation went off the manifold were not relevant, which could further constrain the search for the neural code.

Many successful perturbation studies exploit structural and functional regularities in the system to stay close to the manifold. One property that can help is spatial clustering of neurons representing relevant dimensions of the intrinsic manifold. A salient example is the columnar organization of sensory cortical neurons, such as orientation-selective columns in the primary visual cortex (V1) and direction-selective columns in the middle temporal area (area MT, also known as V5). When neurons within a columnar structure have similar tuning properties, the intrinsic manifold associated with that column becomes approximately one-dimensional. Moreover, because these neurons are spatially clustered, we can control the relevant dimension by increasing or decreasing the overall ensemble activity. The coupling of low dimensionality with spatial clustering allows crude techniques that operate at the level of the column to set the state of the system near its intrinsic manifold. For example, Salzman and Newsome exploited prior knowledge about the clustered nature of direction-tuning neurons in the MT area (Maunsell and Van Essen, 1983; Mikami et al., 1986) to establish a causal role for area MT in motion perception using electrical microstimulation (Salzman et al., 1990). This logic has been used to assess the function of local clusters of neurons in many brain systems in nonhuman primates (Afraz et al., 2006, 2015; Graziano et al., 2002; Klein et al., 2016; Moore and Armstrong, 2003; Robinson, 1972; Robinson and Fuchs, 1969; Romo et al., 1998; Smolyanskaya et al., 2015; Thier and Andersen, 1998; Verhoef et al., 2012) and rodents (Aravanis et al., 2007; Liu et al., 2012; Steinberg et al., 2013; Tye et al., 2011; Witten et al., 2010). We note, however, that, for both electrical microstimulation (Histed et al., 2009, 2013; Tehovnik et al., 2006; Tolias et al., 2005) and optogenetics (Christie et al., 2013; Lin et al., 2005; Mahn et al., 2016; Mattis et al., 2011; Yizhar et al., 2011), the reliability of the interpretations depends on how far the perturbation deviates from what was intended.

Another salient example is the effect of electrical microstimulation and optogenetic activation in V1 (Jazayeri et al., 2012; Tehovnik et al., 2003). Such stimulations evoke patterns of activity that do not match the system's intrinsic manifold, as evidenced by human reports that the stimulation causes an unfamiliar phosphene percept (Foerster, 1929) (in Brindley and Lewin, 1968). Nonetheless, when the stimulation is sufficiently strong, it causes a reflexive orienting behavior, suggesting that the perturbation can penetrate downstream areas despite being off manifold. This suggests that not all dimensions of the intrinsic manifold in V1 are needed to control the animal's orienting behavior.

Many advanced optogenetic experiments aim to maintain perturbations close to the NMOI by targeting specific signaling pathways (Gradinaru et al., 2010; Janak and Tye, 2015; Yizhar et al., 2011). An elegant example comes from a study of corticostriatal processing in auditory discrimination in rats (Znamenskiy and Zador, 2013), where the behavior was influenced by either a non-selective perturbation of the auditory cortex or a selective perturbation of projections from the auditory cortex to the striatum. Remarkably, perturbations of the auditory cortex led to non-selective changes in the animal's performance (e.g., drop of sensitivity), which is analogous to moving off the BMOI. In contrast, targeted corticostriatal projections were able to bias decisions in the direction predicted by the frequency tuning of the stimulated neurons. This experiment demonstrates the importance of avoiding secondary effects to selectively control the behavior.

The case of a negative result (no change in behavior) for an off-manifold perturbation can be more challenging to interpret (Figure 4B, c). Conventional statistics compels us to view the

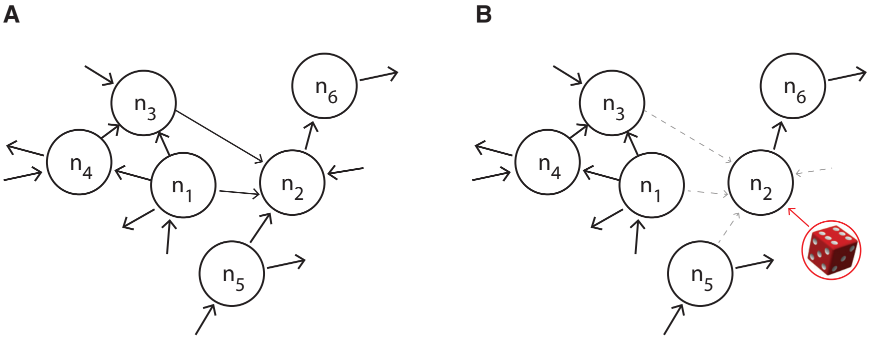

Figure 6. Example of a Belief Network

A belief network is a graphical model composed of a set of nodes. The nodes do not correspond to physical components of the system, such as genes and neurons; instead, they represent propositional variables of interest, some of which we may be able to measure and/or perturb, such as firing rates and behavioral variables. Arrows connect the variables and each arrow represents a presumed causal relationship. We focus on directed acyclic graphs (DAGs) where the network does not contain cyclic motifs. The relationships between any pair of variables is quantified by a conditional probability (i.e., the probability distribution of one variable given knowledge of another variable). Any two nodes that are connected directly are conditionally dependent, and any two nodes that are not connected directly are conditionally independent (although they may be statistically dependent). Using simple probability rules, the network supports the computation of the probabilities of any subset of variables given evidence about any other subset.

(A) A belief network with six variables (n_1–n_6) and their presumed causal interactions (solid arrows).

(B) The same network in a causal experiment in which variable n_2 is experimentally randomized (schematically shown by the red die). This randomization disconnects n_2 from its parent nodes (dashed arrows from n_1, n_3, and n_5), and it creates a reduced graph in which the causal influence of n_2 on n_6 can be examined.

null in a state of suspended disbelief and state conservatively that the experiment fails to reject it (Wagenmakers, 2007). But the nature of this null result is different from an experiment involving an on-manifold perturbation. When we formulate a null hypothesis of the form "variable X does not influence behavior," we implicitly assume that the causal experiment would not put the variable X in a state that the system never visits. Therefore, among the null results, those that arise in the context of off-manifold perturbations are among the least informative. Such observations can be attributed to many different possibilities including the following: (1) the area may be irrelevant for the behavior; (2) the dimensionality of the neural space might be too low to control the behavior; and (3) the perturbation might not be suitably targeting the neural code, even if the targeted neurons contain it.

More generally, off-manifold perturbations demand additional hypotheses and possibilities to be evaluated. For instance, consider a double dissociation experiment in which optogenetic activation of a cell type X (but not Y) causes a change in behavior A (but not B), and vice versa. Under most circumstances, we would take this as strong evidence that cell type X plays a causal role in behavior A and cell type Y in B, and not vice versa. But what if the null effects were because the perturbation was off manifold? What if cell type X is causally involved in behavior B, but the perturbation technique does not mimic the way cell type X functions in that behavior? A recent study in the premotor cortex of mice demonstrates the difficulty of a null effect for off-manifold perturbations (Li et al., 2016). There, it was shown that large-scale unilateral silencing of the premotor cortex had no effect on the selectivity of motor preparatory activity. However, upon more scrutiny, it was found that bilateral silencing strongly disrupted the selectivity of motor preparatory activity. Importantly, this was not simply a matter of silencing a larger volume; bilateral perturbations were more effective even when the volume was accounted for. Evidently, the reason why bilateral inactivation was necessary was that the two hemispheres were tightly coupled and one would compensate for the other. In other words, the intrinsic manifold was constrained by coordinated activity of the two hemispheres, which is presumably why an off-manifold silencing of one hemisphere was not able to alter the behavior. In this case, it was necessary to alter the coordinated activity of both hemispheres to turn the system's cogwheel.

Finally, off-manifold perturbations can engage extraneous pathways that may interfere with the behavior or modulate it in unexpected ways. Therefore, we need to be more sensitive to the possibility of observing non-selective changes in the behavior (Figure 4B, b) or outcomes that are off the intrinsic behavioral manifold (Figure 4B, d and e). An example of how off-manifold perturbations can lead to off-manifold behavioral outcomes is an experiment in which adaptive optics was used to stimulate the retina at a resolution near the width of an individual cone (Hofer et al., 2005). Because the spatial scale of this stimulus is below what the eye sees naturally, one might expect that the cortical circuitry might be blind to it (i.e., no behavioral effect). But single-cone stimulation led to a wide range of color percepts, including some that were never experienced before. While this experiment is extremely valuable in the context of what we currently know about color vision, had we begun to study color perception by randomly perturbing individual cones, we could have conceivably missed the overarching principle of trichromaticity. With this example in mind, we think that off-manifold perturbations are likely to be more useful in dissecting neural systems and behaviors where we have a systematic understanding of the relevant intrinsic manifold.

Intrinsic Manifolds in the Presence of Dynamics

Many behaviors are intrinsically dynamic and, therefore, rely on dynamic patterns of neural activity. Movements with fine temporal structure, cognitive functions such as deliberation and integration (de Lafuente et al., 2015; Hanks et al., 2015; Merchant and Georgopoulos, 2006; Roitman and Shadlen, 2002; Shadlen and Gold, 2004; Thura and Cisek, 2014), and behaviors that depend on knowledge of elapsed time (Eagleman et al., 2005; Finnerty et al., 2015; Janssen and Shadlen, 2005; Jazayeri and Shadlen, 2015; Karmarkar and Buonomano, 2007; Leon and Shadlen, 2003; Merchant et al., 2011, 2013) fall into this category. When the neural code depends on dynamics, causal experiments have to employ perturbations that take those dynamics into account. It is customary to envision dynamics as movements along neural trajectories on a fixed manifold (Churchland et al., 2012; Cunningham and Yu, 2014; Laurent et al., 2001; Mante et al., 2013; Raposo et al., 2014). This is an intuitive representation but it makes the discussion of on- and off-manifold perturbations complicated. Let us consider the evolution of activity during a single trial going from A_1 to A_2 to A_3 to … to A_N. In this representation, points A_i and A_j correspond to two different times during a trial. Let us now consider a perturbation that moves the state from A_3 to A_{10}; i.e., a perturbation that advances the trial through time. Should we consider this perturbation on or off manifold? If we assume dynamics as trajectories on a fixed manifold, then this perturbation should be considered on manifold. However, this is not consistent with our basic definition of intrinsic manifold corresponding to correlational measures with behavior—the behavior never takes a trajectory from A_3 to A_{10}. Similarly, if we consider all points in time as part of the same manifold, some perturbations would correspond to moving back in time (i.e., A_i to A_j with $i > j$), which we would like to consider as off manifold.

To facilitate the assessment of on- versus off-manifold perturbations, we adopt a different representation in which time is added as an independent axis. In this way, the system is associated with one intrinsic manifold at each time point. In this formulation, a fully on-manifold perturbation not only has to respect the instantaneous patterns of neural activity but also should do so over time and within the appropriate timescale.

Creating on-manifold perturbations that respect the dynamics can be remarkably difficult as we do not have suitable techniques to control brain activity over time. It is, therefore, important to ask, in what situations and to what extent do perturbations have to respect the dynamics? Should we treat perturbations that are off manifold with respect to time the same way we treat off-manifold perturbations with respect to the ongoing activity patterns? As we mentioned in our discussion of static intrinsic manifolds, if an off-manifold perturbation is capable of controlling the behavioral variables of interest, we could infer that the neural manifold contains the neural code and we may proceed to search for the relevant features within lower-dimensional state spaces. The analogous idea applies in time: if a perturbation of dynamics at a certain temporal scale controls the behavior, we may not need to analyze behavior at a higher temporal resolution. For example, let us consider an experiment with a trial-based structure in which controlling the average firing rates during the trial controls the behavior selectively. This perturbation is clearly off manifold as it does not respect the higher-order dynamics of firing rates during the trial. However, since it can control the behavior selectively, the experimenter might not need to focus on higher-order dynamics (Histed and Maunsell, 2014). Those dynamics might be extremely important for understanding circuits, biophysics, and possibly other behavioral variables that the experimenter did not measure, but not for understanding the neural code driving the variables on the BMOI. In contrast, if the perturbation fails to control the behavior selectively (Figure 4B, a–c) or leads to unexpected behavioral outcomes (Figure 4B, d and e), the experimenter has to consider the possibility that the reason for failure might be related to the fact that the perturbation did not respect the more fine-grained aspects of the dynamics.

An example of off-manifold perturbation of dynamics that led to a selective control of behavior came from the study of song production in the birdsong. In this system, the temporal precision of the song is attributed to a sparse sequence of bursts in the premotor nucleus higher vocal center (HVC) (Hahnloser et al., 2002; Leonardo and Fee, 2005). Long and Fee devised an elegant perturbation using a cooling device to slow down the underlying dynamics. This perturbation is clearly off manifold with respect to the axis of time but remains on manifold for individual time slices of the neural activity (Long and Fee, 2008). By specifically perturbing the temporal structure of the neural code, they were able to demonstrate that the HVC causally controls the speed of the song without altering its overall profile.

From State Space to Neural Manifold to Neural Code

Consider an experiment aiming to find the neural code for M behavioral variables in a specific behavioral task among N neurons (i.e., an N-dimensional neural state space). Correlational measurements in the absence of any perturbation would characterize the NMOI and BMOI. If we denote the dimensionality of the NMOI and BMOI by η and ψ, respectively, we would expect $\eta \leq N$ and $\psi \leq M$.

After characterizing the correlational correspondence between the NMOI and BMOI (see Figure 3), the experimenter proceeds with using on-manifold perturbations in search of the relevant dimensions that contain the neural code. Let us evaluate two potential outcomes of such perturbations. In one scenario, on-manifold perturbations fail to control the behavior selectively. Since the perturbations are on manifold, this failure means that additional/other neurons must be considered. Note that the same inference cannot be made for off-manifold perturbations. In other words, the failure to control behavior requires the experimenter to increase the dimensionality of the neural state space to increase the probability of targeting the relevant neurons.

In the second scenario, the on-manifold perturbations successfully control the behavior. In this case, the dimensionality of the neural manifold, η, must be at least as large as the dimensionality of the behavioral manifold, ψ. When $\eta = \psi$, the intrinsic manifold contains a compact representation of the neural code. When $\eta > \psi$, it may be possible to continue the search at a lower-dimensional state space. In other words, the success of the on-manifold perturbations invites the experimenter to decrease N.

These two scenarios inform the search for the neural code with respect to the dimensionality of the NMOI (Figure 7). Successful perturbations motivate follow-up experiments on state spaces with reduced dimensionality, and failed perturbations (i.e., ineffective or unselective behavioral outcomes) call for exploration of dimensions not included in the original state space. Note that the choice of the dimensionality of the NMOI may depend on other practical factors. For example, a smaller state space may be preferred as it would facilitate on-manifold perturbations from a technical perspective. This benefit, however, may be offset by reduced selectivity or variance explained with respect to the behavior.

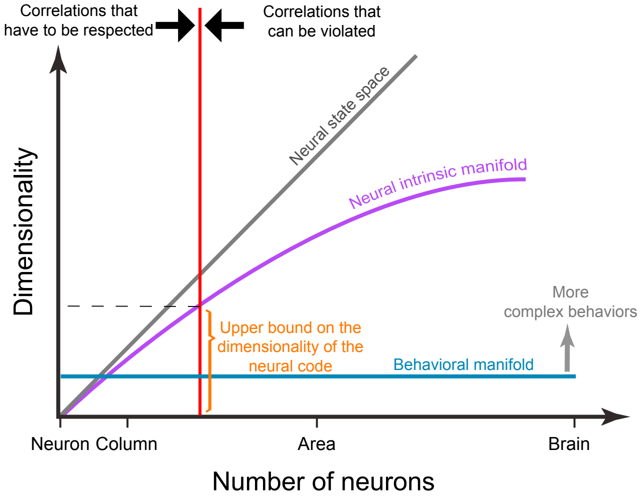
Correlations that have to be respected
Correlations that can be violated
Dimensionality
Neural state space
Neural intrinsic manifold
More complex behaviors
Upper bound on the dimensionality of the neural code
Behavioral manifold
Neuron Column
Area
Brain
Number of neurons

Figure 7. Neurons, Behavior, and the Dimensionality of Neural Codes

The abscissa represents the *number of neurons* included in a recording/perturbation study; in theory, this number can range from a single neuron to all neurons in the brain. The scale bar is schematic and does not depict the actual scale magnification within this large range. The ordinate represents *dimensionality* and is used to show the dimensionality of both neural and behavioral measures. The gray unity line indicates that the dimensionality of the neural state space is the same as the number of neurons (by definition). The purple line shows that the dimensionality of the neural intrinsic manifold also increases with the number of neurons. The upper bound of this dimensionality is set by the neural state space (gray line), but, in practice, it can be much lower (see text). The horizontal blue line shows the dimensionality of the behavioral manifold. This dimensionality is independent of the number of neurons, and it is instead dictated by the behavioral paradigm and the number of behavioral variables that are measured experimentally. Dimensionality of the behavioral manifold is typically far lower than that of the neural state space, although using complex behaviors and more behavioral variables can increase this number (gray upward arrow). The vertical red line represents the smallest state space in which on-manifold perturbations can control the behavior selectively. Note that on-manifold perturbations at this scale are off manifold in state spaces that include other neurons (Figure 2C). The position of the red line may vary depending on the behavioral manifold of interest. It may shifts to the far left (small state spaces), if only a few neurons could selectively control the behavior, or to the far right (larger state spaces), if the behavior is controlled by the coordinated activity of many neurons in the brain. The crossing point between the red line and the purple line represents the minimum dimension with respect to the neural manifold of interest, which is an upper bound for the dimensionality of the key features (i.e., neural code) (orange bracket).

In thinking about the relevant dimensions of analysis, it may be useful to develop benchmarks for the dimensions of intrinsic manifolds and the neural code with respect to the state space. Naturally, the primary determinant of the lower bound on N is the dimensionality of the behavioral manifold ψ. However, in reality, N is usually much larger than ψ for two reasons. First, many behaviors depend on converging and redundant signals from large populations of similarly behaving neurons (Cohen and Maunsell, 2009; Froudarakis et al., 2014; Jazayeri and Movshon, 2006; Priebe and Lisberger, 2004; Rust et al., 2006; Shadlen et al., 1996; Zohary et al., 1994). Second, neurons in the state space are usually correlated (i.e., due to interactions and/or common input), such that η is much smaller than N.

Note that, our representation of intrinsic manifold implicitly assumes that a single variable (e.g., firing rate) captures what each neuron encodes. If it were the case that a neuron could convey information in more than one way, in effect, acting like multiple single-variable neurons, then the dimensionality of the state space would become larger than the number of neurons. However, the dimensionality of the NMOI, η, would still be bounded from below by ψ.

As a final point, we note that the lowest dimensional neural intrinsic manifold that contains a neural code may still have more dimensions than needed to capture the behavioral variables. In other words, it is possible that multiple neural states would correspond to the same measured behavioral variables. Many factors could contribute to such phenomenon. For example, neural activity might control aspects of behavior (e.g., prior expectations) that are not measured or included on the BMOI (e.g., choice behavior). Similarly, neural activity might be modulated by nuisance variables (e.g., contrast of a stimulus) that is not relevant for a certain behavioral paradigm and, thus, not accounted for by the experimenter. The presence of such factors renders the smallest η that contains a neural code an upper bound for the dimensionality of the neural code (Figure 7).

Concluding Remarks

Our aim was to provide a critical assessment of how correlational and causal approaches advance our understanding of the neural codes that link the brain to behavior. Noting that the full state space is enormous, we proposed that the search space can be shrunk by looking at intrinsic manifolds (i.e., subspaces that are characterized by correlations). Additionally, we used the analogy of the cogwheel to argue that perturbations can powerfully expose structure and logic of the neural code when (1) they interface with the system at an appropriate scale and (2) they respect the intrinsic patterns of activity at the scale at which they are applied. This process might need new tools that go beyond targeting individual genes, cell types, and local ensembles of neurons. The next generation of tools we envision would be able to randomize correlated patterns of activity and navigate the state space through coordinated perturbations that may span multiple elements of the circuit simultaneously.

ACKNOWLEDGMENTS

We thank Troy Littleton, Greg Horwitz, J. Anthony Movshon, and James DiCarlo for comments on the manuscript and the lab members of Mehrdad Jazayeri, James DiCarlo, and Nancy Kanwisher for thoughtful discussions. M.J. is supported by the NIH (NINDS-NS078127), the Sloan Foundation, the Klingenstein Foundation, the Simons Foundation, the Center for Sensorimotor Neural Engineering, and the McGovern Institute.

REFERENCES

Adamala, K.P., Martin-Alarcon, D.A., and Boyden, E.S. (2016). Programmable RNA-binding protein composed of repeats of a single modular unit. Proc. Natl. Acad. Sci. USA *113*, E2579–E2588.

Afraz, S.-R., Kiani, R., and Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. Nature *442*, 692–695.

Afraz, A., Boyden, E.S., and DiCarlo, J.J. (2015). Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination. Proc. Natl. Acad. Sci. USA *112*, 6730–6735.

Aravanis, A.M., Wang, L.-P., Zhang, F., Meltzer, L.A., Mogri, M.Z., Schneider, M.B., and Deisseroth, K. (2007). An optical neural interface: in vivo control of rodent motor cortex with integrated fiberoptic and optogenetic technology. J. Neural Eng. *4*, S143–S156.

Brindley, G.S., and Lewin, W.S. (1968). The sensations produced by electrical stimulation of the visual cortex. J. Physiol. *196*, 479–493.

Brody, C.D., Hernández, A., Zainos, A., and Romo, R. (2003). Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex. Cereb. Cortex *13*, 1196–1207.

Buschman, T.J., and Miller, E.K. (2007). Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. Science *315*, 1860–1862.

Christie, I.N., Wells, J.A., Southern, P., Marina, N., Kasparov, S., Gourine, A.V., and Lythgoe, M.F. (2013). fMRI response to blue light delivery in the naïve brain: implications for combined optogenetic fMRI studies. Neuroimage *66*, 634–641.

Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Foster, J.D., Nuyujukian, P., Ryu, S.I., and Shenoy, K.V. (2012). Neural population dynamics during reaching. Nature *487*, 51–56.

Cohen, M.R., and Maunsell, J.H.R. (2009). Attention improves performance primarily by reducing interneuronal correlations. Nat. Neurosci. *12*, 1594–1600.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., and Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Cunningham, J.P., and Yu, B.M. (2014). Dimensionality reduction for large-scale neural recordings. Nat. Neurosci. *17*, 1500–1509.

de Lafuente, V., Jazayeri, M., and Shadlen, M.N. (2015). Representation of accumulating evidence for a decision in two parietal areas. J. Neurosci. *35*, 4306–4318.

DiCarlo, J.J., and Cox, D.D. (2007). Untangling invariant object recognition. Trends Cogn. Sci. *11*, 333–341.

Eagleman, D.M., Tse, P.U., Buonomano, D., Janssen, P., Nobre, A.C., and Holcombe, A.O. (2005). Time and the brain: how subjective time relates to neural time. J. Neurosci. *25*, 10369–10371.

Finnerty, G.T., Shadlen, M.N., Jazayeri, M., Nobre, A.C., and Buonomano, D.V. (2015). Time in Cortical Circuits. J. Neurosci. *35*, 13912–13916.

Fitzgerald, J.K., Freedman, D.J., Fanini, A., Bennur, S., Gold, J.I., and Assad, J.A. (2013). Biased associative representations in parietal cortex. Neuron *77*, 180–191.

Foerster, O. (1929). Beitrage zur pathophysiologie der sehbahn und der spehsphare. J. Psychol. Neurol. *39*, 435–463.

Froudarakis, E., Berens, P., Ecker, A.S., Cotton, R.J., Sinz, F.H., Yatsenko, D., Saggau, P., Bethge, M., and Tolias, A.S. (2014). Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. Nat. Neurosci. *17*, 851–857.

Ganguli, S., Bisley, J.W., Roitman, J.D., Shadlen, M.N., Goldberg, M.E., and Miller, K.D. (2008). One-dimensional dynamics of attention and decision making in LIP. Neuron *58*, 15–25.

Georgopoulos, A.P. (1994). Population activity in the control of movement. Int. Rev. Neurobiol. *37*, 103–119, discussion 121–123.

Gradinaru, V., Zhang, F., Ramakrishnan, C., Mattis, J., Prakash, R., Diester, I., Goshen, I., Thompson, K.R., and Deisseroth, K. (2010). Molecular and cellular approaches for diversifying and extending optogenetics. Cell *141*, 154–165.

Graziano, M.S.A., Taylor, C.S.R., and Moore, T. (2002). Complex movements evoked by microstimulation of precentral cortex. Neuron *34*, 841–851.

Hahnloser, R.H.R., Kozhevnikov, A.A., and Fee, M.S. (2002). An ultra-sparse code underlies the generation of neural sequences in a songbird. Nature *419*, 65–70.

Hanks, T.D., Kopec, C.D., Brunton, B.W., Duan, C.A., Erlich, J.C., and Brody, C.D. (2015). Distinct relationships of parietal and prefrontal cortices to evidence accumulation. Nature *520*, 220–223.

Harvey, C.D., Coen, P., and Tank, D.W. (2012). Choice-specific sequences in parietal cortex during a virtual-navigation decision task. Nature *484*, 62–68.

Histed, M.H., and Maunsell, J.H.R. (2014). Cortical neural populations can guide behavior by integrating inputs linearly, independent of synchrony. Proc. Natl. Acad. Sci. USA *111*, E178–E187.

Histed, M.H., Bonin, V., and Reid, R.C. (2009). Direct activation of sparse, distributed populations of cortical neurons by electrical microstimulation. Neuron *63*, 508–522.

Histed, M.H., Ni, A.M., and Maunsell, J.H.R. (2013). Insights into cortical mechanisms of behavior from microstimulation experiments. Prog. Neurobiol. *103*, 115–130.

Hofer, H., Singer, B., and Williams, D.R. (2005). Different sensations from cones with the same photopigment. J. Vis. *5*, 444–454.

Holdefer, R.N., and Miller, L.E. (2002). Primary motor cortical neurons encode functional muscle synergies. Exp. Brain Res. *146*, 233–243.

Janak, P.H., and Tye, K.M. (2015). From circuits to behaviour in the amygdala. Nature *517*, 284–292.

Janssen, P., and Shadlen, M.N. (2005). A representation of the hazard rate of elapsed time in macaque area LIP. Nat. Neurosci. *8*, 234–241.

Jazayeri, M., and Movshon, J.A. (2006). Optimal representation of sensory information by neural populations. Nat. Neurosci. *9*, 690–696.

Jazayeri, M., and Shadlen, M.N. (2015). A Neural Mechanism for Sensing and Reproducing a Time Interval. Curr. Biol. *25*, 2599–2609.

Jazayeri, M., Lindbloom-Brown, Z., and Horwitz, G.D. (2012). Saccadic eye movements evoked by optogenetic activation of primate V1. Nat. Neurosci. *15*, 1368–1370.

Johnson, K.O. (2000). Neural coding. Neuron *26*, 563–566.

Karmarkar, U.R., and Buonomano, D.V. (2007). Timing in the absence of clocks: encoding time in neural network states. Neuron *53*, 427–438.

Kaufman, M.T., Churchland, M.M., Ryu, S.I., and Shenoy, K.V. (2014). Cortical activity in the null space: permitting preparation without movement. Nat. Neurosci. *17*, 440–448.

Klapoetke, N.C., Murata, Y., Kim, S.S., Pulver, S.R., Birdsey-Benson, A., Cho, Y.K., Morimoto, T.K., Chuong, A.S., Carpenter, E.J., Tian, Z., et al. (2014). Independent optical excitation of distinct neural populations. Nat. Methods *11*, 338–346.

Klein, C., Evrard, H.C., Shapcott, K.A., Haverkamp, S., Logothetis, N.K., and Schmid, M.C. (2016). Cell-Targeted Optogenetics and Electrical Microstimulation Reveal the Primate Koniocellular Projection to Supra-granular Visual Cortex. Neuron *90*, 143–151.

Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C.E., Kepecs, A., Mainen, Z.F., Qi, X.-L., Romo, R., Uchida, N., and Machens, C.K. (2016). Demixed principal component analysis of neural population data. eLife *5*, e10989.

Laurent, G., Stopfer, M., Friedrich, R.W., Rabinovich, M.I., Volkovskii, A., and Abarbanel, H.D. (2001). Odor encoding as an active, dynamical process: experiments, computation, and theory. Annu. Rev. Neurosci. *24*, 263–297.

Leon, M.I., and Shadlen, M.N. (2003). Representation of time by neurons in the posterior parietal cortex of the macaque. Neuron *38*, 317–327.

Leonardo, A., and Fee, M.S. (2005). Ensemble coding of vocal control in birdsong. J. Neurosci. *25*, 652–661.

Li, N., Daie, K., Svoboda, K., and Druckmann, S. (2016). Robust neuronal dynamics in premotor cortex during motor planning. Nature *532*, 459–464.

Lin, L., Osan, R., Shoham, S., Jin, W., Zuo, W., and Tsien, J.Z. (2005). Identification of network-level coding units for real-time representation of episodic experiences in the hippocampus. Proc. Natl. Acad. Sci. USA *102*, 6125–6130.

Liu, X., Ramirez, S., Pang, P.T., Puryear, C.B., Govindarajan, A., Deisseroth, K., and Tonegawa, S. (2012). Optogenetic stimulation of a hippocampal engram activates fear memory recall. Nature *484*, 381–385.

Lo, C.-C., and Wang, X.-J. (2006). Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. Nat. Neurosci. *9*, 956–963.

Long, M.A., and Fee, M.S. (2008). Using temperature to analyse temporal dynamics in the songbird motor pathway. Nature *456*, 189–194.

Ma, W.J., Beck, J.M., Latham, P.E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nat. Neurosci. *9*, 1432–1438.

Mahn, M., Prigge, M., Ron, S., Levy, R., and Yizhar, O. (2016). Biophysical constraints of optogenetic inhibition at presynaptic terminals. Nat. Neurosci. *19*, 554–556.

Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature *503*, 78–84.

Marshel, J.H., Mori, T., Nielsen, K.J., and Callaway, E.M. (2010). Targeting single neuronal networks for gene expression and cell labeling in vivo. Neuron *67*, 562–574.

Mattis, J., Tye, K.M., Ferenczi, E.A., Ramakrishnan, C., O'Shea, D.J., Prakash, R., Gunaydin, L.A., Hyun, M., Fenno, L.E., Gradinaru, V., et al. (2011). Principles for applying optogenetic tools derived from direct comparative analysis of microbial opsins. Nat. Methods *9*, 159–172.

Maunsell, J.H., and Van Essen, D.C. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. J. Neurophysiol. *49*, 1127–1147.

Merchant, H., and Georgopoulos, A.P. (2006). Neurophysiology of perceptual and motor aspects of interception. J. Neurophysiol. *95*, 1–13.

Merchant, H., Zarco, W., Pérez, O., Prado, L., and Bartolo, R. (2011). Measuring time with different neural chronometers during a synchronization-continuation task. Proc. Natl. Acad. Sci. USA *108*, 19784–19789.

Merchant, H., Harrington, D.L., and Meck, W.H. (2013). Neural basis of the perception and estimation of time. Annu. Rev. Neurosci. *36*, 313–336.

Mikami, A., Newsome, W.T., and Wurtz, R.H. (1986). Motion selectivity in macaque visual cortex. I. Mechanisms of direction and speed selectivity in extrastriate area MT. J. Neurophysiol. *55*, 1308–1327.

Moore, T., and Armstrong, K.M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. Nature *421*, 370–373.

Otchy, T.M., Wolff, S.B.E., Rhee, J.Y., Pehlevan, C., Kawai, R., Kempf, A., Gobes, S.M.H., and Ölveczky, B.P. (2015). Acute off-target effects of neural circuit manipulations. Nature *528*, 358–363.

Penfield, W., and Jasper, H. (1954). Epilepsy and the Functional Anatomy of the Human Brain (Little, Brown and Company).

Priebe, N.J., and Lisberger, S.G. (2004). Estimating target speed from the population response in visual area MT. J. Neurosci. *24*, 1907–1916.

Prigge, M., Schneider, F., Tsunoda, S.P., Shilyansky, C., Wietek, J., Deisseroth, K., and Hegemann, P. (2012). Color-tuned channelrhodopsins for multiwavelength optogenetics. J. Biol. Chem. *287*, 31804–31812.

Raposo, D., Kaufman, M.T., and Churchland, A.K. (2014). A category-free neural population supports evolving demands during decision-making. Nat. Neurosci. *17*, 1784–1792.

Rigotti, M., Barak, O., Warden, M.R., Wang, X.-J., Daw, N.D., Miller, E.K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature *497*, 585–590.

Robinson, D.A. (1972). Eye movements evoked by collicular stimulation in the alert monkey. Vision Res. *12*, 1795–1808.

Robinson, D.A., and Fuchs, A.F. (1969). Eye movements evoked by stimulation of frontal eye fields. J. Neurophysiol. *32*, 637–648.

Roitman, J.D., and Shadlen, M.N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. J. Neurosci. *22*, 9475–9489.

Romo, R., Hernández, A., Zainos, A., and Salinas, E. (1998). Somatosensory discrimination based on cortical microstimulation. Nature *392*, 387–390.

Rust, N.C., Mante, V., Simoncelli, E.P., and Movshon, J.A. (2006). How MT cells analyze the motion of visual patterns. Nat. Neurosci. *9*, 1421–1431.

Sadtler, P.T., Quick, K.M., Golub, M.D., Chase, S.M., Ryu, S.I., Tyler-Kabara, E.C., Yu, B.M., and Batista, A.P. (2014). Neural constraints on learning. Nature *512*, 423–426.

Salzman, C.D., Britten, K.H., and Newsome, W.T. (1990). Cortical microstimulation influences perceptual judgements of motion direction. Nature *346*, 174–177.

Sanes, J.R., and Lichtman, J.W. (1999). Can molecules explain long-term potentiation? Nat. Neurosci. *2*, 597–604.

Shadlen, M.N., and Gold, J.I. (2004). The neurophysiology of decision making as a window on cognition. In The Cognitive Neurosciences, Third Edition, M.S. Gazzaniga, ed. (MIT Press), pp. 1229–1241.

Shadlen, M.N., Britten, K.H., Newsome, W.T., and Movshon, J.A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. J. Neurosci. *16*, 1486–1510.

Siegel, M., Buschman, T.J., and Miller, E.K. (2015). Cortical information flow during flexible sensorimotor decisions. Science *348*, 1352–1355.

Smolyanskaya, A., Haefner, R.M., Lomber, S.G., and Born, R.T. (2015). A Modality-Specific Feedforward Component of Choice-Related Activity in MT. Neuron *87*, 208–219.

Steinberg, E.E., Keiflin, R., Boivin, J.R., Witten, I.B., Deisseroth, K., and Janak, P.H. (2013). A causal link between prediction errors, dopamine neurons and learning. Nat. Neurosci. *16*, 966–973.

Tehovnik, E.J., Slocum, W.M., and Schiller, P.H. (2003). Saccadic eye movements evoked by microstimulation of striate cortex. Eur. J. Neurosci. *17*, 870–878.

Tehovnik, E.J., Tolias, A.S., Sultan, F., Slocum, W.M., and Logothetis, N.K. (2006). Direct and indirect activation of cortical neurons by electrical microstimulation. J. Neurophysiol. *96*, 512–521.

Thier, P., and Andersen, R.A. (1998). Electrical microstimulation distinguishes distinct saccade-related areas in the posterior parietal cortex. J. Neurophysiol. *80*, 1713–1735.

Thura, D., and Cisek, P. (2014). Deliberation and commitment in the premotor and primary motor cortex during dynamic decision making. Neuron *81*, 1401–1416.

Tolias, A.S., Sultan, F., Augath, M., Oeltermann, A., Tehovnik, E.J., Schiller, P.H., and Logothetis, N.K. (2005). Mapping cortical activity elicited with electrical microstimulation using FMRI in the macaque. Neuron *48*, 901–911.

Tye, K.M., Prakash, R., Kim, S.-Y., Fenno, L.E., Grosenick, L., Zarabi, H., Thompson, K.R., Gradinaru, V., Ramakrishnan, C., and Deisseroth, K. (2011). Amygdala circuitry mediating reversible and bidirectional control of anxiety. Nature *471*, 358–362.

Verhoef, B.-E., Vogels, R., and Janssen, P. (2012). Inferotemporal cortex subserves three-dimensional structure categorization. Neuron *73*, 171–182.

Vogelstein, J.T., Park, Y., Ohyama, T., Kerr, R.A., Truman, J.W., Priebe, C.E., and Zlatic, M. (2014). Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning. Science *344*, 386–392.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. Psychon. Bull. Rev. *14*, 779–804.

Weiskrantz, L. (1968). Analysis of Behavioral Change (Harper & Row).

Witten, I.B., Lin, S.-C., Brodsky, M., Prakash, R., Diester, I., Anikeeva, P., Gradinaru, V., Ramakrishnan, C., and Deisseroth, K. (2010). Cholinergic interneurons control local circuit activity and cocaine conditioning. Science *330*, 1677–1681.

Yizhar, O., Fenno, L.E., Davidson, T.J., Mogri, M., and Deisseroth, K. (2011). Optogenetics in neural systems. Neuron *71*, 9–34.

Znamenskiy, P., and Zador, A.M. (2013). Corticostriatal neurons in auditory cortex drive decisions during auditory discrimination. Nature *497*, 482–485.

Zohary, E., Shadlen, M.N., and Newsome, W.T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. Nature *370*, 140–143.