Journal
Perspective

How the Brain Might Work: Statistics Flowing in Redundant Population Codes

Xaq Pitkow[1,2] and Dora E Angelaki[1,2]
[1] Department of Neuroscience, Baylor College of Medicine
[2] Department of Electrical and Computer Engineering, Rice University

We propose that the brain performs approximate probabilistic inference using nonlinear recurrent processing in redundant population codes. Different overlapping patterns of neural population activity encode the brain's estimates and uncertainties about latent variables that could explain its sense data. Nonlinear processing implicitly passes messages about these variables along a graph that determines which latent variables interact according to an internal model of the world. Since there are many equivalent neural implementations of this computation, we describe a general approach to identify the essential features of the neural algorithm. This approach uses dimensionality reduction in redundant codes to extract from the fine-grained neural signals how task-relevant variables are represented and transformed. To reveal these fundamental computations, it is insufficient to record neural activity during simple tasks because such tasks do not probe the brain's structured internal model. Instead, core inferential brain functions can only be revealed by studying large-scale activity patterns during moderately complex, naturalistic behaviors.

INTRODUCTION

Perception as inference

In its purest form, probabilistic inference is the 'right' way to solve problems (Laplace 1812). While animal brains face various constraints and cannot always solve problems in the smartest possible way, many human and animal behaviors do provide strong evidence of probabilistic computation (Heeger and Simoncelli 1993; Gallistel et al. 2001; Ernst and Banks 2002; Körding and Wolpert 2004; Yang and Shadlen 2007; Cheng et al. 2007). The idea that the brain performs statistical inference harkens back at least to Helmholtz (1925). According to this hypothesis, the goal of sensory processing is to identify properties of the world based on ambiguous sensory evidence. Since the true properties cannot be determined with perfect confidence, there is a probability distribution associated with different interpretations, and animals weigh these probabilities when choosing actions.

This idea has an illustrious history, with too many contributors for a complete list: (Barlow 1969; Hinton and Sejnowski 1993; Knill and Richards 1996; Lee and Mumford 2003; Friston 2010; Yuille and Kersten 2006; Stocker and Simoncelli 2009; Knill and Pouget 2004; Rao 2004; Doya et al. 2007; Denève 2008; Hoyer and Hyvärinen 2003; Berkes et al. 2011; Yu and Dayan 2005; Tenenbaum et al. 2011). But despite excellent work interpreting behavior as probabilistic inference, and many models and experiments relating neuronal activity to probabilities, there is yet no consensus about how the brain actually implements these models.

To understand the neural basis of the brain's probabilistic computations, we need to understand the overlapping processes of encoding, recoding, and decoding. *Encoding* describes the relationship between sensory stimuli and neural activity patterns. *Recoding* describes the dynamic transformation of those patterns into other patterns. *Decoding* describes the use of neural activity to generate actions.

In this paper we speculate how these processes might work, and what can be done to test it. Our core hypothesis is that the brain uses nonlinear computation by redundant, recurrently-connected population codes to perform statistical inference. This hypothesis is deliberately general-purpose and abstract, but it should be tested in concrete cases. We posit that to reveal the structure of these computations, we must study large-scale activity patterns in the brains of animals performing naturalistic tasks of greater complexity than most current efforts. We also argue that since there are many equivalent ways for the brain to implement natural computations, one can understand them best at the representational level — characterizing how encoded *task variables* are affected by neural computations — rather than by fine details of how the large-scale neural activity patterns are transformed.

But before we address inference in complex tasks, we'll discuss some things neuroscience has successfully learned from simple tasks, and where those tasks necessarily fail to reveal computational principles.

Insights from simple tasks, and remaining questions

Neuroscience has learned much in the past several decades using a simple kind of tasks in which subjects choose between two options — two-alternative forced-choice tasks (2AFC). Many scientific advances came from measuring how populations of neurons encode and decode information about

simple stimuli. Binary tasks made it easier to isolate specific features and to measure perceptual or behavioral differences definitively, with less data. Based on such measurements, computational studies began asking how much of the encoded information was successfully decoded.

Many experiments found individual neurons that were tuned to task stimuli with enough reliability that only a small handful could be averaged together to perform as well or better than the animal in a 2AFC task (Newsome et al. 1989; Cohen & Newsome 2009; Gu et al. 2008; Chen et al. 2013a). With a brain full of neurons, why isn't behavior better? One possible answer is that responses to a fixed stimulus are correlated, and these covariations cannot be averaged away by pooling. They thus limit the information the animal has about the world, creating a redundant neural code (Zohary et al. 1994, Moreno-Bote et al. 2014). After properly accounting for correlated noise, is neural processing "optimal", reaching the intrinsic bounds on behavioral performance, or "suboptimal", falling far short of this bound? The answer may depend on the task, but the question itself has spurred valuable discussion and controversy in the interpretation of experiments.

Amazingly, individual neural responses also covary with choices or reported percepts, even when the stimulus itself is perfectly ambiguous, with no task-relevant information and thus no correct answer. Correlations between neural responses and choices ('choice correlations') have been reported in multiple tasks and cortical areas (Britten et al. 1996; Uka & DeAngelis 2004; Nienborg & Cumming 2007; Gu et al. 2008; Fetsch et al. 2011; Chen et al. 2013a,b,c; Liu et al. 2013a,b). The origin of these choice correlations remains unresolved. Do they reflect inherited 'bottom-up' noise, whose accumulation forms the perceptual decision (Britten et al. 1996; Shadlen and Newsome 2001; Gold and Shadlen 2003; Parker and Newsome 1998; Schall 2003; Yang and Shadlen 2007; Shadlen et al. 1996; Pitkow et al. 2015)? Or are these choice correlations caused by useful top-down signals related to feature attention, high-level prior, or some computationally useful internal state (Krug 2004; Nienborg & Cumming 2007, 2009, 2010; Nienborg et al. 2012; Berkes et al. 2011; Reimer et al. 2014; Orbán et al. 2016; Haefner et al. 2016)?

Causal manipulations seem more reliable than correlation for assessing the role of neural circuits, but interpreting such experiments is still complicated. Inactivation of areas with high choice correlations sometimes produce large behavioral deficits (Chowdhury and DeAngelis 2008) — but not always (Chen et al. 2016; Katz et al. 2016). Furthermore, it is commonly argued that if inactivating a brain area produces no behavioral deficit, then it does not contribute to the behavior. Yet that is only guaranteed to be true for optimal computation: one may observe no performance change if the natural circuitry overweights the area as much as it is underweighted after inactivation. Even when there is an effect, it may be hard to interpret because stimuli and behavior are often represented in multiple interconnected areas, and inactivating one may change the responses of others in complex ways. Since basic questions about distributed processing remains, it will be critical to record from multiple brain areas during simple tasks.

Overall, through studies of how a neuron's activity correlates with simple stimuli, simple behaviors, and other neurons, these 2AFC tasks have enabled insights into how sensory evidence is represented and accumulated for perceptual decisions.

A critique of simple tasks

However, such simple tasks do create other problems. The most fundamental is that they limit the computations and neural activity to a domain where the true power and adaptability of the brain is hidden. When the tasks are low-dimensional, the mean neural population dynamics are bound to a low-dimensional subspace, and measured neural activity seems to hit this bound (Gao and Ganguli 2015). This means that the low-dimensional responses observed in the brain may be an artifact of overly simple tasks. Even worse, many of our standard tasks are linearly solvable using trivial transformations of sense data. And if natural tasks could be solved with linear computation, then we wouldn't need a brain! We could just wire our sensors to our muscles and accomplish the same goal, because multiple linear processing steps is equivalent to a single linear processing step. Distinguishing these steps becomes extremely difficult at best, and uninterpretable at worst.

Finally, principles that govern neural computation in overtrained animals performing unnatural lab tasks may be not generalize. Are we learning about the real brain in action, or a laboratory artifact? Evolution did not optimize brains for 2AFC tasks, and the real benefit of complex inferences like weighing uncertainty may not be apparent unless the uncertainty has complex structure. How can we understand how the brain works without challenging it with the tasks for which it evolved?

ALGORITHM OF THE BRAIN

The challenge of perception

In perception, the quantities of interest — the things we can act upon — cannot be directly observed through our senses. These unobservable quantities are called latent or hidden variables. For example, when we reach for a mug, we never directly sense the object's three-dimensional boundary — that is latent — but only receive stereo images of reflected light, and an increase in tactile pressure when our joint angles reach some value. Some latent quantities are relevant to behavioral goals, like the handle's orientation, while other latent variables are a nuisance, like shadows of other objects. Perception is hard because both types of latent variables affect sensory observations, and we must disentangle nuisance variables from our sense data to isolate the task-relevant ones (DiCarlo and Cox 2007).

We must infer all of this based on uncertain sensory evidence. There are multiple sources of uncertainty. Some is intrinsic to physics: lossy observations due to occlusion or photon shot noise. Some is unresolvable variation, like the hum of a city street. Other uncertainty is due to biology, including neural noise and limited sampling by the sensors and subsequent computation. Uncertainty also arises from suboptimal processing (Beck et al. 2012): model mismatch

behaves much like structured noise. Regardless of its origin, since uncertainty is an inevitable property of perceptual systems, it is valuable to process signals in accordance with probabilistic reasoning.

Unfortunately, exact probabilistic inference is intractable for models that are as complex as the ones our brains seem to make. First, merely representing arbitrary joint probabilities exactly requires enormous resources, exponential in the number of variables. Second, performing inference over these distributions requires an exponentially large number of operations. This means that exact inference in arbitrary models is out of the question, for the brain or any other type of computer. Finally, even exploiting the structure in the natural world, a lifetime of experience never really has enough data to constrain a complete statistical model, nor do we have enough computational power and time to perform statistical inference based on these ideal statistics. Our brain must invoke the 'blessing of abstraction' (Goodman et al. 2009) to overcome this 'curse of dimensionality' (Bellman 1957). The brain must make assumptions about the world that limit what it can usefully represent, manipulate and learn — this is the 'no free lunch theorem' (Wolpert 1996).

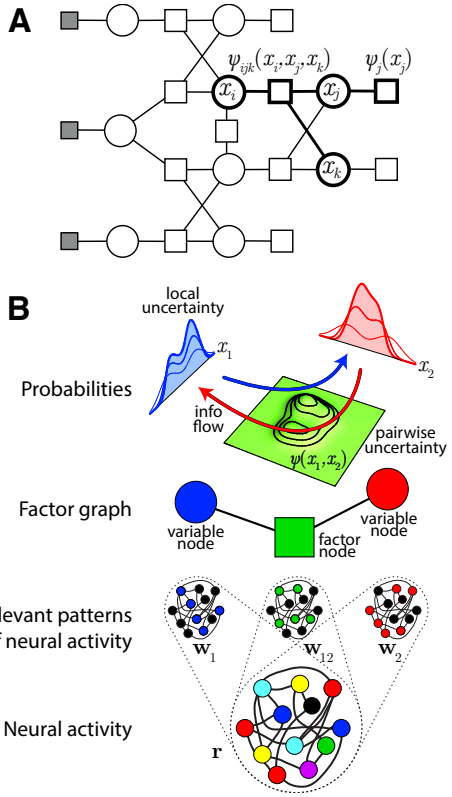

Figure 1. Inference in graphical models embedded in neural activity. **A**. A probabilistic graphical model represents direct relationships (conditional dependencies) amongst many variables. Here this is depicted as a factor graph, with circles showing variables x_i and squares showing interactions ψ between subsets of variables. Three variable nodes and two interaction nodes are highlighted. Here we depict a graphical model which does not naturally describe causality, but this can be readily generalized to allow one-way interactions across time. **B**. Illustration of our neural inference model. Statistics that encode probabilities over latent variables are low-dimensional properties embedded in high-dimensional spatiotemporal neural activity patterns. As neuronal activities evolve over time, these patterns exchange information along a sparse interaction graph that models the joint distribution over latent variables. The neural dynamics thereby represent the dynamics of probabilistic inference.

Encoding: Redundant distributed representations of probabilistic graphical models

Since the probability distribution over things in the world is hugely complex, we hypothesize that the brain simplifies the world by assuming that not every variable necessarily interacts with all other variables. Instead, there may be a small number of important interactions. Variables and their interactions can be elegantly visualized as a sparsely connected graph (Figure 1A), and described mathematically as a probabilistic graphical model (Koller and Friedman 2009). These are representations of complex probability distributions as products of lower-dimensional functions (see Box). Such constraints on possible distributions are appropriate for the natural world, which has both hierarchical and local structures.

Knowledge about the world is embodied in these interactions between variables. Many of the most important ones express nonlinear relationships between variables. For

Probabilistic inference and Population codes

Probabilistic inference: Drawing conclusions based on ambiguous observations. Typical inference problems include finding the marginal probability of a task-relevant variable, or finding the most probable explanation of observed data.

Probabilistic computation: Transformation of signals in a manner consistent with rules of probability and statistics, especially through appropriate sensitivity to uncertainty (Ma 2012).

Latent variables (also called hidden or causal variables): quantities whose value cannot be directly observed, yet determine observations. Latent variables may be *task-relevant* or *irrelevant* (*nuisance)*, depending on the task.

Probabilistic graphical model: a decomposition of a probability distribution as a product of functions that describe interactions between subsets of variables. One useful such model is a factor graph (Figure 1A) that represents a structured probability distribution $P(\mathbf{x}_\alpha) = \prod_\alpha \psi_\alpha(\mathbf{x}_\alpha)$ where $\mathbf{x}=(x_1,\ldots,x_n)$ is a vector of all variables and $\mathbf{x}_\alpha$ is a subset of variables that interact through the function, or factor, $\psi_\alpha(\mathbf{x}_\alpha)$.

Statistical interaction: dependency between two variables that cannot be explained by other observed covariates. This may be generated by real causal interactions in the world, or due to some unobserved latent variables.

Higher-order interaction: A nonlinear statistical interaction between variables. An especially interesting case is when three or more variables interact. This leads naturally to contextual gating, whereby one variable (the 'context') determines whether other variables interact. See (Ranzato and Hinton 2010) for example.

Message-passing algorithm: An iterative sequence of computations that performs a global computation by operating locally on statistical information ('messages') conveyed along a probabilistic graphical model.

Population code: Representation of a sensory, motor, or latent variable by the collective activity of many neurons. We can estimate the information content of a population by optimal decoding.

Information-limiting correlations (informally, 'bad noise'): Covarying noise fluctuations in large populations that are indistinguishable from changes in the encoded variable. These arise when sensory signals are embedded in a higher-dimensional space, or when suboptimal upstream processing throws away extensive amounts of information. These noise correlations cannot be averaged away by adding more neurons (Moreno-Bote et al. 2014).

Redundancy: An extreme degeneracy when different signals are perfectly interchangeable. If two neuronal populations inherit the same limited information from an upstream source, then either population can be decoded separately, or the two can be averaged, and the result is the same.

Robustness: Insensitivity to variations in network weights. A computation is robust whenever uncertainty added by suboptimal processing is much smaller than the intrinsic uncertainty caused by information-limiting noise.

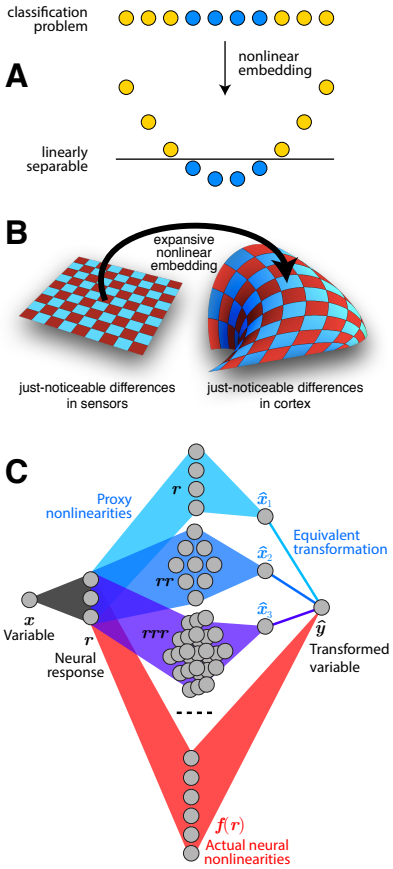

Figure 2. Nonlinearity and redundancy in neural computation

A: The classification task of separating yellow from blue cannot be accomplished by linear operations, because the task-relevant variable (color) is entangled with a nuisance variable (horizontal position). After embedding the data nonlinearly into a higher-dimensional space, the task-relevant variable becomes linearly separable. **B**: Signals from the sensory periphery have limited information content, illustrated here by a cartoon tiling of an abstract stimulus space. Each square tile represents just-noticeable differences between responses, i.e. the uncertainty or limited resolution at which the responses can be reliably discriminated due to limited sampling and noise. When sensory signals are transformed and embedded into a higher-dimensional space, the uncertainty is processed exactly the same way. This produces a redundant population code with information-limiting high-order correlations (Moreno-Bote et al. 2014, Yang and Pitkow 2015), leading to just-noticeable differences between cortical responses that are tilted in the cortical response space. **C**: In such a redundant code, the precise nonlinear transformations of individual neurons (red) are just one of many ways to accomplish a given nonlinear transformation of the encoded variable. Since the fine details do not matter, it is valuable to model the more abstract level where the nonlinearity affects the information content. For this purpose, simple nonlinearities (e.g. polynomials) may be a convenient representation. However, the most natural nonlinearities to examine are those statistics that are tuned to the task-relevant variables. In complex tasks, these statistics may be complicated functions of the neural responses, so many layers of nonlinearity may be warranted.

example, unlike pervasive models of sparse image coding (Olshausen and Field 1997), natural images are not generated as a sum of images. Instead many statistical features of natural images arise from the occlusion of objects (Pitkow 2010) and the multiplicative absorption of light (Wainwright and Simoncelli 2000).

How are these probabilistic graphical models represented by neural activity? Information about each sensory variable is spread across spiking activity of many neurons with similar stimulus sensitivities. Conversely, neurons are also tuned to many different features of the world (Rigotti et al. 2013). Together, these facts mean that the brain uses a distributed and multiplexed code.

There are several competing models of how neurons encode probabilities. In spatial representations of probability, the encoded probability distribution is determined by the spatial pattern of which neurons are active (Ma et al. 2006; Jazayeri and Movshon 2006; Savin and Denève 2014; Rao 2004). For example, in linear probabilistic population codes, every neural spike adds log-probability to some interpretations of a scene, so more spikes typically means more confidence (Ma et al. 2006; Jazayeri and Movshon 2006).

In temporal representations of probability, such as the sampling hypothesis (Hoyer and Hyvärinen 2003) instantaneous neural activity represents a single interpretation, without uncertainty. Probabilities are reflected instead by the set of interpretations over time (Hoyer and Hyvärinen 2003; Berkes et al. 2011; Moreno-Bote et al. 2011; Buesing et al. 2011; Haefner et al. 2016; Orbán et al. 2016). These models have different advantages and disadvantages, and account for different aspects of experimental observations.

Neurons can also participate in both spatial and temporal patterns: a temporal code over feature amplitudes could also serve as a spatial code over task variables, depending on how the neural signals are decoded. In a similar vein, neurons can participate in multiple spatial codes, with different projections of population activity encoding biases and precisions (Ma et al. 2006; Ma 2010; Beck et al. 2011) about distinct variables (Figure 1B, Denève et al. 2007; Beck et al. 2012; Raju and Pitkow 2016).

It is also possible to have a spatiotemporal representation that encodes some aspects of uncertainty in both spatial patterns and their changes over time (Lee and Mumford 2003; Savin and Denève 2014). For example, a temporal mixture of probabilistic population codes could use spatial encoding for precise modes, and fluctuate over time to allow for multiple interpretations.

Any of these representations will have a limit to the information it encodes, whether about estimates of latent variables or a distribution over them. This information limit may arise at the sensory input itself, or may be worsened by biological constraints or suboptimal computation (Beck et al. 2012; Babadi and Sompolinsky 2014). Once signals enter the brain, the neural representation expands massively, engaging many times more neurons than sensory receptors. Despite the large increase in the number of neurons, the brain cannot encode more information than it receives, so all of these extra neurons can at best recode the relevant signals in a new form (Figure 2B), and may lose information (Babadi and Sompolinsky 2014). The result is that cortical codes are highly redundant (Zohary et al. 1994; Moreno-Bote et al. 2014; Pitkow et al. 2015), essentially possessing many copies of the same information in different groups of neurons. Below we explain how the resultant redundancy has a major impact on how we should think about and describe neural computation.

Recoding: Inference by message-passing

Natural tasks require nonlinear computation, whether probabilistic or not, because most nuisance variables are entangled with task-relevant variables (DiCarlo and Cox 2007). Figure 2A shows a simple example that illustrates how nonlinear computation can allow subsequent linear computation to perform well. The total postsynaptic input to a neuron can be well described as linear sums of the presynaptic activities. This means that upstream nonlinearities allow such a neuron to implement complex classifications.

Ethological tasks require far more complex nonlinearities to untangle task-relevant properties from nuisance variables. In principle, any untangling can be accomplished by a simple network with one layer of nonlinearities, since this architecture is a universal function approximator—both for feedforward nets (Cybenko 1989; Hornik 1991) and recurrent nets (Schäfer and Zimmerman 2007). However, in practice this can be more easily accomplished by a 'deep' cascade of simpler nonlinear transformations. This may be because the parameters are easier to learn, because the hierarchical structure imposed by the deep model are better matched to natural inputs (Montúfar et al. 2014), because certain representations use brain resources more economically, or all of the above. Indeed, trained artificial deep neural networks have notable similarities with biological neural networks (Yamins et al. 2014).

To what extent are these nonlinear networks probabilistic? One can trivially interpret neural responses as encoding probabilities, because neuronal responses differ upon repeated presentations of the same stimulus and according to Bayes' rule that means that any given neural response could arise from multiple different stimuli. But to actually *use* that encoding to perform probabilistic inference (exactly or approximately), the brain must transform its information in a manner that accounts for trial-by-trial changes in uncertainty (Ma 2012). Thus the relevance of encoded probabilities is inextricably linked to their use. For this reason, the choice of task is critical to understand recoding: one can only test neural models of probabilistic inference in tasks where uncertainty varies over time and this variation affects behavior.

Exact inference in probabilistic models is generally intractable except in special cases. Many algorithms for inference in probabilistic graphical models are based on transmitting information about probability distributions along the graph of interactions. These algorithms go by the name of 'message-passing' algorithms, because the information they convey between nodes can be viewed as messages. This broad class of algorithms includes belief propagation (Pearl 1988), expectation propagation (Minka 2001), mean-field inference, and other types of variational inference (Wainwright and Jordan 2008). Even some forms of sampling (Geman and Geman 1984; Lee and Mumford 2003) can be viewed as message-passing algorithms with a random component. Each algorithm is defined by how incoming information is combined and how outgoing information is selected. The differences between algorithms reflect different choices of local approximations for intractable global computations. For instance, belief propagation treats all incoming messages as independent, even if they are not. Some randomness may be useful to represent and compute with distributions (Hinton and Sejnowski 1983; Hoyer and Hyvärinen 2003) as well as overcome blind spots in suboptimal message-passing algorithms (Pitkow et al. 2011). Amongst the diverse possibilities, the commonality is that recurrent nonlinear transformations disseminate statistics along a graph that reflects direct interactions between latent variables.

Mathematically speaking, message-passing algorithms operate on these statistical summaries of latent variables. But in any practical implementation on a computer, the algorithms operate on binary strings that represent the underlying variables. The best way to understand the computation is not to examine the transformation of individual bits, but to look instead at the transformation of the variables those bits encode. Likewise, in the brain, we propose that it is more fundamental to describe the nonlinear transformation of encoded variables than to describe the detailed nonlinear response properties of individual neurons (Figure 1B) (Kriegeskorte et al. 2008; Yang and Pitkow 2015; Raju and Pitkow 2016), although the two nonlinearities can be related. Since neural network computations can implement computationally useful transformations in multiple ways, we should therefore focus on the shared properties of equivalent computations. This abstracts away the fine implementation details while preserving the essential properties of the nonlinear computation (see Box, Figure 2C). This is a valid and quantifiable abstraction in redundant codes (Figure 2B, Pitkow et al. 2014; Yang and Pitkow 2015).

Although we extoll the virtues of abstracting away from individual neuronal nonlinearities, nonetheless there may be certain functions that are difficult to implement as a combination of generic nonlinearities. For instance, both a quadratic nonlinearity and divisive normalization can be implemented as a sum of sigmoidally transformed inputs, but the latter requires a much larger number of neurons (Raju and Pitkow 2015). We speculate that cell types are hard-wired with specialized connectivity (Kim et al. 2014; Jiang et al. 2015) in order to accomplish useful operations, like divisive normalization, that are harder to learn by adjusting synapses between arbitrarily connected neurons.

Decoding: Probabilistic control

If an animal never guides any action by task-relevant information encoded by its neural populations, then it doesn't matter that neurons encode that information, or even if the network transforms it the right way. Thus it is critical to measure how the neural representations relate to behavior. Ideally, we would like to predict variations in behavior from fluctuations in neural activity.

Choosing a good action can be formulated as a control problem, where the animal aims to maximize expected utility, i.e. to get the best long-term subjective benefit while weighing uncertainty. Maximizing utility involves building not only a model of the external world, but also a model of the animal's causal influence on it.

Because the world is uncertain, an animal may choose actions that have a low probability of gaining a reward directly, but which gather enough predictive information to increase the reward probability in the future (Sutton and Barto 1998; Bialek et al. 2001). This tradeoff between exploration and exploitation is a key element of such natural behaviors as foraging (Charnov 1976, Stephens and Krebs 1986).

The brain appears to use multiple strategies to map its beliefs onto actions, depending on the task and the animal's ability to model the task structure. Multiple brain areas and neurotransmitters have been implicated in both learning and using these strategies (Sutton and Barto 1998; Rao 2010; Yu and Dayan 2005). Probabilistic inference appears to play a major role in guiding action, which would make our theory of statistics flowing through redundant population codes especially useful for understanding computation all the way from encoding, through recoding, to decoding.

From Theory to Experiment

If this theory of neural computation is correct, how could we test it? This general-purpose computation can best be revealed in concrete naturalistic tasks with interesting interactions. However, truly natural stimuli are too complex, and too filled with uncontrolled and indescribable nuisance variations, to make good computational theories (Rust and Movshon 2005). On the other hand, things should be made "as simple as possible, but no simpler" (Prausnitz 2002). We want to understand the remarkable properties of the brain — especially those aspects that still go far beyond the piecewise-linear fitting of impressively successful deep networks (Krizhevsky et al. 2012). This means we need to challenge it to be flexible, to adjust processing dynamically. This requires us to find a happy medium: tasks that are neither too easy nor too hard.

To reveal the brain's internal model, good experimental tasks must require predictions — actions that cannot be based on current evidence but on extrapolations into the future. Prediction is hard, especially about the future (Shapiro 2006). Ergo, prediction tasks typically involves significant uncertainty, which gives us the opportunity to measure the neural substrate of probabilistic inference.

To make this work, we need some fairly big data. Such data is now becoming accessible. Large-scale recording technology allows us to monitor up to a thousand neurons at once (Stevenson and Körding 2011). Efforts are underway to record from a million neurons simultaneously sometime in the next decade (Alivisatos et al. 2015). Chronic recordings give us more data to judge long-term shifts in redundant codes (Sadtler et al. 2014). Biomarkers such as pupilometry and motion tracking provide additional observations that we know influence neural activity (Reimer et al. 2014). Wireless transmission and neurologging allow us to observe the brain activity of untethered animals, who are then freer to pursue more natural behavioral strategies.

How should we analyze all of this rich data to better understand the brain?

We expect that much of the brain's machinery is dedicated to attributing dynamic latent causes to observations, and using them to choose appropriate actions. To understand this process, we need some experimental handle on the latent variables we expect to see. Inferring latent variables requires perceptual models, and we measure an animal's percept through its behavior, so we need behavioral models. We call attention to two types here. The first is a black-box model, such as an artificial recurrent neural network trained on a task. One can compare then the structure of the artificial network activity to the structure of real brain activity. If representational similarity (Kriegeskorte et al. 2008) between them suggests that the solutions are similar, one can then analyze the fully-observable artificial machinery to gain some insights into neural computation. This approach has been used fruitfully by (Mante et al. 2013; Yamins et al. 2014). However, such models are difficult to interpret without some external guess about the relevant latent variables and how they influence each other. In simple tasks, the relevant latent variables may be intuitively obvious. In complex tasks we may not know how to interpret the computation beyond the similarity to the artificial network (Yamins et al. 2014), which makes it hard to understand and generalize. Ultimately, we need some principled way to characterize the latent variables.

This leads us to the second type of behavioral model: optimal control. Such a model uses probabilistic inference to identify the state of time-varying latent variables, their interactions, and the actions that maximize expected value. Clearly, animals are not universally optimal. Nonetheless, for some tasks, animals may understand the structure of the task while mis-estimating its parameters. Consequently, we can use the optimal structure to direct our search for computational features in neural networks.

Figure 3 shows a schematic for how one can use a behavioral model with identifiable latent variables to interpret neural activity. Step 1 is to find the *encoding*, that is, the distributed neural representations of the latent variables. In the context of statistical inference, these representations ought to include not only guesses about the true state, but also uncertainties about those states. This encoding provides us with a substantial dimensionality reduction, allowing us to abstract away many fine details about neural encoding and concentrate on the information content. Based on that dimensionality reduction, we can then predict the latent variables in new neural recordings that were not used to find the encoding. Step 2 is to measure the interactions between the estimates and uncertainties about latent variables, which determines the process of *recoding*. Statistical inference via message-passing defines how these quantities should interact, and can generate strong predictions. The brain may have learned clever tricks for this inference, which we can measure experimentally by computing the interactions between the brain's internal estimates — at least, those we estimated from neural data (yes, estimates of estimates). Step 3 is to use brain activity and our model of interactions to predict actions, i.e. to predict *decoding*. This will be particularly revealing during periods of greatest uncertainty, since this is when the sensory stimuli are weakest and the animal's internal model will be most

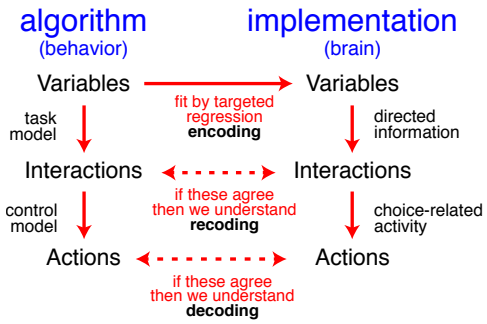

Figure 3. Schematic for understanding distributed codes

A behavioral model describes task-relevant latent variables, how they interact statistically, and the algorithm by which they guide actions. A corresponding model of the brain quantifies the neural encoding of those variables, the interactions between them, and how they relate to behavior. A good match between the behavioral and brain models provides evidence that these neural recordings reflect the encoding, recoding and decoding processes. A poor match implies either that our behavioral model is wrong such that the brain has found other latent variables that explain its observations, or that we are not recording from populations that encode the predicted latent variables and mediate their interactions. We can then revise the behavioral model based on these observations, and either record from other brain areas or simplify the task to focus on latent variables that are well represented in the recorded areas.

valuable. This is a generalization of the choice correlations in simple tasks that we described above.

In essence, this analysis framework allows one to work at the representational level, a step removed from the neural mechanisms, to measure the encoding, recoding, and decoding algorithm of the brain.

Task design to reveal flexible probabilistic computation

To understand flexible brain computations, the tasks we present to an animal should satisfy certain requirements. First, to understand nonlinear computations, one should include nuisance variables that the brain must untangle. Second, to reveal probabilistic inference, which hinges on appropriate treatment of uncertainty, one must manipulate uncertainty experimentally. Third, to expose an animal's internal model, the task should require the animal to predict the future, for otherwise the animal can rely upon visible evidence which can compensate for any false beliefs an animal might harbor. A task based on prediction also makes it simpler to identify neuronal fluctuations that relate directly to behavior and not to the input. Fourth, the task should be naturalistic, but neither too easy nor too hard. This has the best chances of keeping the brain engaged and the animal incentivized.

Based on these considerations, we commend foraging as an excellent candidate task. In foraging, an animal searches for rewards based on sensory cues in an uncertain environment, and it can take distinct actions to acquire either rewards or information. Uncertainty about reward encourages an exploration-exploitation tradeoff that provides evidence about the animal's predictions and thus about its internal model. Foraging is naturalistic and allows flexible animal behavior, while allowing experimenters to control the contingencies that indicate value. It includes components of perceptual decision-making and thus leverages our existing knowledge about neural circuits. These virtues would address the problems arising from overly simple tasks and would help us refine our understanding of the neural basis of behavior.

Conclusion

In this paper, we proposed that the brain naturally performs probabilistic inference, and critiqued overly simple tasks as being ill suited to expose the inferential computations that make the brain special. We introduced a hypothesis about computation in cortical circuits. Our hypothesis has three parts. First, overlapping patterns of population activity encode statistics that summarize both estimates and uncertainties about latent variables. Second, the brain specifies how those variables are related through a sparse probabilistic graphical model of the world. Third, recurrent circuitry implements a nonlinear message-passing algorithm that selects and localizes the brain's statistical summaries of latent variables, so that all task-relevant information is actionable.

We also suggested experiments that could provide evidence about this hypothesis. These experiments should be based on naturalistic tasks that require the animal to predict uncertain future rewards, and thereby reveal its internal model of the environment. By recording from many neurons across multiple brain areas, and relating the activity to the stimulus, predicted latent variables, and actions, we can analyze the neuronal interactions that constitute neural computation.

Finally, we emphasized the advantage in studying the computation at the level of neural population activity, rather than at the level of single neurons or membrane potentials: If the brain does use redundant population codes, then many fine details of neural processing don't matter for computation. Instead it can be beneficial to characterize computation at a more abstract level, operating on variables encoded by populations, rather than on the substrate.

ACKNOWLEDGEMENTS

The authors thank Jeff Beck, Greg DeAngelis, Ralf Haefner, Kaushik Lakshminarasimhan, Ankit Patel, Alexandre Pouget, Paul Schrater, Andreas Tolias, Rajkumar Vasudeva Raju, and Qianli Yang for valuable discussions. XP was supported in part by a grant from the McNair Foundation, NSF CAREER Award IOS-1552868, a generous gift from Britton Sanderford, and by the Intelligence Advanced Research Projects Activity (IARPA)

via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003.[1] XP and DA are supported by the Simons Collaboration on the Global Brain award 324143, and BRAIN Initiative grants NIH 5U01NS094368-02 and NSF 1450923 BRAIN 43092-N1.

REFERENCES

Alivisatos AP, Miyoung Chun M, Church GM, Greenspan RJ, Roukes ML, Yuste R (2015). A National Network of Neurotechnology Centers for the BRAIN Initiative, *Neuron* **88**(3): 445–8.

Babadi B, Sompolinsky H (2014). Sparseness and expansion in sensory representations. *Neuron*. 83(5): 1213–26.

Barlow HB (1969). Pattern recognition and the responses of sensory neurons. *Annals of the New York Academy of Sciences*. 156(2): 872–81.

Beck JM, Latham PE, Pouget A (2011). Marginalization in neural circuits with divisive normalization. *J neurosci*. 31(43): 15310–9.

Beck J*, Ma WJ*, Pitkow X, Latham P, Pouget A (2012). Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron* 74(1): 30–9.

Beck J, Pouget A, Heller KA (2012). Complex inference in neural circuits with probabilistic population codes and topic models. *Advances in neural information processing systems*.

Bellman RE (1957). *Dynamic Programming*. Princeton University Press.

Berkes P, Orbán G, Lengyel M, Fiser J (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* 331: 83–87.

Bialek W, Nemenman I, Tishby N (2001). Predictability, complexity, and learning. *Neural computation*. 13(11):2409–63.

Britten KH, Newsome WT, Shadlen MN, Celebrini S, Movshon JA (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis Neurosci*. 13: 87-100.

Buesing L, Bill J, Nessler B, Maass W (2011). Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Comp Bio*. 7(11): e1002211.

Charnov EL (1976). Optimal foraging, the marginal value theorem. *Theor Popul Biol*. 9: 129-136.

Chen A, Deangelis GC, Angelaki DE (2013a). Functional specializations of the ventral intraparietal area for multisensory heading discrimination. *J Neurosci*. 33: 3567–3581.

Chen X, Deangelis GC, Angelaki DE (2013b). Diverse spatial reference frames of vestibular signals in parietal cortex. *Neuron*. 80: 1310–1321.

Chen X, DeAngelis GC, Angelaki DE (2013c). Eye-centered representation of optic flow tuning in the ventral intraparietal area. *J Neurosci*. 33: 18574–18582.

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*. 36(13): 3789–98.

Cheng K, Shettleworth SJ, Huttenlocher J, Rieser JJ (2007). Bayesian integration of spatial information. *Psychological bulletin*. 133(4): 625–37.

Chowdhury SA, DeAngelis GC (2008). Fine discrimination training alters the causal contribution of macaque area MT to depth perception. *Neuron*. 60: 367–377.

Cohen MR, Newsome WT (2009). Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. *J Neurosci*. 29(20): 6635–48.

Cybenko G (1989). Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems. 2(4):303-14.

Denève S (2008). Bayesian spiking neurons I: inference. *Neural computation*. 20(1): 91–117.

Denève S, Duhamel JR, Pouget A (2007). Optimal sensorimotor integration in recurrent cortical networks: a neural implementation of Kalman filters. *J Neurosci*. 27(21): 5744–56.

DiCarlo JJ, Cox DD (2007). Untangling invariant object recognition. *Trends in cognitive sciences*. 11(8): 333–41.

Doya K, Ishii S, Pouget A, Rao R (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT press.

Ernst MO, Banks MS (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*. 415(6870): 429–33.

Gao P, Ganguli S (2015). On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology* 32:148-55.

Geman S, Geman D (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*. 6: 721–41.

Gold JI, Shadlen MN (2003). The influence of behavioral context on the representation of a perceptual decision in developing oculomotor commands. The *J Neurosci*. 23(2): 632–51.

Goodman ND, Ullman TD, Tenenbaum JB (2009). Learning a theory of causality. Psychological review. 118(1):110–9.

Gu Y, Angelaki DE, DeAngelis GC (2008). Neural correlates of multisensory cue integration in macaque MSTd. *Nat neurosci*. 11(10):1201–10.

Fetsch CR, Pouget A, DeAngelis GC, Angelaki DE (2011). Neural correlates of reliability-based cue weighting during multisensory integration. *Nat Neurosci*. 15: 146–154.

Friston K (2010). The free-energy principle: a unified brain theory?. *Nature Reviews Neuroscience*. 11(2): 127–38.

Gallistel CR, Mark TA, King AP, Latham PE (2001). The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. *J Exp Psych: Animal Behavior Processes*. 27(4):354–72.

Haefner R, Berkes P, Fiser J (2016). Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. *Neuron* 90(3): 649–60.

Heeger DJ, Simoncelli EP (1993). Model of visual motion sensing. *Spatial vision in humans and robots*. 19: 367–92.

Helmholtz H (1925). *Physiological optics* III. Optical Society of America: 318.

Hinton GE, Sejnowski TJ (1983). Optimal perceptual inference. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 448–453.

Hornik K (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*. 4(2): 251–7.

Hoyer PO, Hyvärinen A (2003). Interpreting Neural Response Variability as Monte Carlo Sampling of the Posterior. *Advances in Neural Information Processing Systems*.

Jazayeri M, Movshon JA (2006). Optimal representation of sensory information by neural populations. *Nat neurosci*. 9(5):690–6.

Jiang X, Shen S, Cadwell CR, Berens P, Sinz F, Ecker AS, Patel S, Tolias AS. (2015) Principles of connectivity among morphologically defined cell types in adult neocortex. *Science*. 350(6264): aac9462.

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

Kim JS, Greene MJ, Zlateski A, Lee K, Richardson M, Turaga SC, Purcaro M, Balkam M, Robinson A, Behabadi BF, Campos M, Denk W, Seung S, EyeWirers (2014). Space-time wiring specificity supports direction selectivity in the retina. *Nature*. 509(7500): 331–6.

Knill DC, Pouget A (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*. 27(12): 712–9.

Knill DC, Richards W (1996). *Perception as Bayesian inference*. Cambridge University Press.

Koller D, Friedman N (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Körding KP, Wolpert DM (2004). Bayesian integration in sensorimotor learning. *Nature*. 427(6971): 244–7.

Krizhevsky A, Sutskever I, Hinton GE (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*.

[1] The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government.

Kriegeskorte N, Mur M, Bandettini PA (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in systems neuroscience* 2:4.

Krug K (2004). A common neuronal code for perceptual processes in visual cortex? Comparing choice and attentional correlates in V5/MT. *Phil Trans Roy Soc London B*. 359: 929–941.

Laplace PS (1812). *Théorie Analytique des Probabilités* (Ve Courcier, Paris).

Lee TS, Mumford D (2003). Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A*. 20(7):1434–48.

Liu S, Dickman JD, Newlands SD, DeAngelis GC, Angelaki DE (2013a). Reduced choice-related activity and correlated noise accompany perceptual deficits following unilateral vestibular lesion. *PNAS*. 110: 17999–18004.

Liu S, Gu Y, DeAngelis GC, Angelaki DE (2013b). Choice-related activity and correlated noise in subcortical vestibular neurons. *Nat Neurosci*. 16: 89–97.

Ma WJ*, Beck J*, Latham P, Pouget A (2006). Bayesian inference with probabilistic population codes. *Nat Neurosci* 9(11): 1432–8.

Ma WJ (2010). Signal detection theory, uncertainty, and Poisson-like population codes. *Vision Research* 50: 2308–19.

Ma WJ (2012). Organizing probabilistic models of perception. *Trends in cognitive sciences*. 16(10):511-8.

Mante V, Sussillo D, Shenoy KV, Newsome WT (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*. 503(7474):78-84.

Minka TP (2001). Expectation propagation for approximate Bayesian inference. Uncertainty in Artificial Intelligence (UAI), ed. Breese and Koller, Morgan Kaufmann Pub. 362–369.

Montúfar GF, Pascanu R, Cho K, Bengio Y (2014). On the number of linear regions of deep neural networks. *Advances in Neural Information Processing Systems*.

Moreno-Bote M, Knill D, Pouget A (2011). Bayesian sampling in visual perception. PNAS 108(30): 12491–6.

Moreno-Bote R, Beck J, Kanitscheider I, Pitkow X, Latham P, Pouget A (2014). Information-limiting correlations. *Nat neurosci*. 17(10): 1410–7.

Newsome WT, Britten KH, Movshon JA (1989). Neuronal correlates of a perceptual decision. Nature 341: 52–4.

Nienborg H, Cohen MR, Cumming BG (2012). Decision-related activity in sensory neurons: correlations among neurons and with behavior. *Ann Rev Neurosci*. 35: 463–483.

Nienborg H, Cumming BG (2007). Psychophysically measured task strategy for disparity discrimination is reflected in V2 neurons. *Nat Neurosci*. 10: 1608–1614.

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

Nienborg H, Cumming B (2010). Correlations between the activity of sensory neurons and behavior: how much do they tell us about a neuron's causality? *Curr Opin Neurobiol*. 20: 376–381.

Olshausen BA, Field DJ (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*. 381(6583): 607–9.

Orbán G, Berkes P, Fiser J, Lengyel M (2016). Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*.

Parker AJ, Newsome WT (1998). Sense and the single neuron: probing the physiology of perception. *Annual rev neuroscience*. 21(1): 227–77.

Pearl J (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Pitkow X (2010). Exact feature probabilities in images with occlusion. *Journal of Vision*. 10(14):42, 1–20.

Pitkow X, Ahmadian Y, Miller KD (2011). Learning unbelievable probabilities. *Advances in Neural Information Processing Systems*. 738–746.

Pitkow X, Lakshminarasimhan K, Pouget A (2014). How brain areas can predict behavior, yet have no influence. *COSYNE Abstract*.

Pitkow X, Liu S, Angelaki DE, DeAngelis GC, Pouget A (2015). How can single sensory neurons predict behavior? *Neuron*. 87(2): 411–23.

Prausnitz F (2002). *Roger Sessions: How a "Difficult" Composer Got That Way*. Oxford University Press. (paraphrasing Einstein)

Ranzato M, Hinton G (2010). Modeling Pixel Means and Covariances Using Factorized Third-Order Boltzmann Machines. CVPR.

Rao RP (2004). Bayesian computation in recurrent neural circuits. *Neural computation*. 16(1):1-38.

Rao RP (2010). Decision making under uncertainty: a neural model based on partially observable markov decision processes. *Frontiers in computational neuroscience*. 4:146.

Raju R, Pitkow X (2015). Marginalization in random nonlinear neural networks. *American Physical Society Meeting Abstracts*. Vol 1.

Raju R, Pitkow X (2016). Inference by Reparameterization in Neural Population Codes. *Advances in Neural Information Processing Systems*.

Reimer J, Froudarakis E, Cadwell CR, Yatsenko D, Denfield GH, Tolias AS (2014). Pupil fluctuations track fast switching of cortical states during quiet wakefulness. *Neuron*. 84(2):355-62.

Rigotti M, Barak O, Warden MR, Wang X-J, Daw ND, Miller EK, Fusi S (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature* 497: 585–590.

Rust NC, Movshon JA (2005). In praise of artifice. *Nat. Neurosci*. 8(12):1647-50.

Sadtler PT, Quick KM, Golub MD, Chase SM, Ryu SI, Tyler-Kabara EC, Byron MY, Batista AP (2014). Neural constraints on learning. *Nature*. 512(7515):423-6.

Savin C, Denève S (2014). Spatio-temporal representations of uncertainty in spiking neural networks. *Advances in Neural Information Processing Systems* 27.

Schäfer AM, Zimmerman HG (2007). Recurrent neural networks are universal approximators. *Int. J of Neural Systems*, 17(4): 253–263.

Schall JD (2003). Neural correlates of decision processes: neural and mental chronometry. *Current opinion in neurobiology*. 13(2): 182–6.

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(4): 1486–510.

Shadlen MN, Newsome WT (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J Neurophysiol*. 86(4): 1916–36.

Shapiro FR (2006). *The Yale Book of Quotations*. Yale University Press.

Stephens DW, Krebs JR (1986). *Foraging Theory*. Princeton University Press.

Stevenson IH, Kording KP (2011). How advances in neural recording affect data analysis. *Nat neurosci*. 14(2):139–42.

Stocker A, Simoncelli EP (2009). A Bayesian model of conditioned perception. *Advances in neural information processing systems*. 1409–1416.

Sutton RS, Barto AG (1998). *Reinforcement learning: An introduction*. Cambridge: MIT press, 1998.

Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*. 331(6022): 1279–85.

Uka T, DeAngelis GC (2004). Contribution of area MT to stereoscopic depth perception: choice-related response modulations reflect task strategy. *Neuron*. 42: 297–310.

Wainwright MJ, Jordan MI (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*. 1(1–2): 1–305.

Wainwright MJ, Simoncelli EP (2000). Scale mixtures of Gaussians and the statistics of natural images. *Advances in neural information processing systems*. 12(1): 855–861.

Wolpert D (1996). The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation* **8**: 1341–1390.

Yamins D*, Hong H*, Cadieu C, Solomon EA, Seibert D, DiCarlo JJ (2014). Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex. *PNAS*. doi: 10.1073/pnas.1403112111.

Yang Q, Pitkow X (2015). Robust nonlinear neural codes. *COSYNE Abstract*.

Yang T, Shadlen MN (2007). Probabilistic reasoning by neurons. *Nature*. 447(7148): 1075–80.

Yu AJ, Dayan P (2005). Uncertainty, neuromodulation, and attention. *Neuron*. 46(4): 681–92.

Yuille A, Kersten D (2006). Vision as Bayesian inference: analysis by synthesis?. *Trends in cognitive sciences*. 10(7): 301–8.

Zohary E, Shadlen MN, Newsome WT (1994). Correlated neuronal discharge rate and its implication for psychophysical performance. *Nature* 370: 140–143.