Bayesian encoding and decoding as distinct perspectives on neural coding

Richard D. Lange, Sabyasachi Shivkumar, Ankani Chattoraj, & Ralf M. Haefner

The Bayesian brain hypothesis is one of the most influential ideas in neuroscience. However, unstated differences in how Bayesian ideas are operationalized make it difficult to draw general conclusions about how Bayesian computations map onto neural circuits. Here, we identify one such unstated difference: some theories ask how neural circuits could recover information about the world from sensory neural activity (Bayesian decoding), whereas others ask how neural circuits could implement inference in an internal model (Bayesian encoding). These two approaches require profoundly different assumptions and lead to different interpretations of empirical data. We contrast them in terms of motivations, empirical support and relationship to neural data. We also use a simple model to argue that encoding and decoding models are complementary rather than competing. Appreciating the distinction between Bayesian encoding and Bayesian decoding will help to organize future work and enable stronger empirical tests about the nature of inference in the brain.

According to the Bayesian brain hypothesis, neural circuits carry out statistical computations by combining prior knowledge with new evidence, combining multiple sources of information according to their reliability, and taking actions that account for uncertainty. In the case of perception, prior knowledge is assumed either to come from experience or to be encoded genetically, having been learned over the course of evolution. While any given sensory measurement may be noisy or ambiguous—providing a wide likelihood function in Bayesian terms—prior knowledge is deployed to resolve these ambiguities when possible.

We characterize and contrast two distinct perspectives on the Bayesian brain hypothesis, which we call ‘Bayesian encoding’ and ‘Bayesian decoding’. These are complementary perspectives that make different assumptions about the nature of the inference problems faced by the brain, and are supported or falsified by different kinds of empirical data. We argue that not making their differences explicit has led to confusion about how to interpret empirical data. For example, the debate on whether neural responses are better modeled as samples or parameters of probability distributions is complicated by the fact that sampling codes usually make assumptions consistent with Bayesian encoding, whereas parametric codes often make assumptions consistent with Bayesian decoding.

We illustrate the complementarity using a simple model of primary visual cortex (V1) in which we construct a sampling-based encoding model based on a simple model of natural images, and derive the implied decoding model from it, finding that the decoding model is a kind of canonical parametric code known as a probabilistic population code (PPC). We conclude with a discussion of distributional neural codes in general.

Bayesian encoding versus Bayesian decoding

We follow the seminal work of Zemel et al. in assuming that patterns of neural activity represent entire probability distributions over a variable, not just a point estimate of it, that is, that they form a distributional
Bayesian encoding

We define Bayesian encoding as the view that the primary function of sensory neurons is to compute and represent an approximation to some predefined probability distribution over relevant variables. We use the term ‘encoding’ because the distribution that neurons are hypothesized to represent conceptually precedes the actual neural responses. That is, in Bayesian encoding models, there exists a reference distribution that is defined independently of how neurons actually respond, and which is approximately encoded by neural responses.

Bayesian encoding requires a source for the reference distribution. In the context of the sensory system, this typically takes the form of an internal generative model of sensory inputs, and the distribution to be encoded is the posterior over latent variables in that model (Fig. 1a,b). With this perspective, the goal of sensory areas of the brain is to learn a statistical model of its sensory inputs in which sensory observations, such as an image on the retina, are explained as the result of higher-order causes or latent variables, such as objects, lights and textures. A generative model makes this process explicit by assigning prior probabilities to the (co)occurrence of latent variables and by quantifying the likelihood of a particular configuration of the causes for generating a particular sensory observation. The encoded posterior distribution in this framework is defined over the latent variables in this statistical model.

For latent variables \( x \) and sensory input \( I \), optimal inference means computing the posterior distribution,

\[
p_b(x|I) = \frac{p_b(I|x)p_b(x)}{p_b(I)}.
\]

We use the subscript letter ‘b’ in \( p_b(\cdot) \) to refer to quantities in the brain’s internal model, and to distinguish them from other types of probability such as a decoder’s uncertainty. Bayesian encoding poses the question of how neural circuits could compute and represent the posterior distribution \( p_b(x|I) \) for any sensory input \( I \) (Fig. 1c). In general, exact inference is an intractable problem, leading to the question of how the brain could compute and represent an approximation to the true posterior (Fig. 1d–f), and what the nature of this approximation is. This line of reasoning motivates work on ‘neurally plausible approximate inference algorithms,’ including approaches with connections to sampling-based inference (Fig. 1e), as well as parametric neural codes (Fig. 1d; reviewed in refs. 4,17,18).

Bayesian decoding

We define Bayesian decoding as the perspective in which neural activity is treated as given, and emphasis is placed on the statistical uncertainty of a decoder observing this neural activity. Any stimulus \( s \) that elicits neural responses \( r \) is optimally decoded by computing \( p(s|r) \) (Fig. 2), and so Bayesian decoding is closely related to ideal observer models in psychophysics, often involving tasks that require the estimation of scalar aspects of some sensory input (for example, its orientation or its contrast). Note that we distinguish between the ‘stimulus’ \( s \) in a task (for example, orientation) and the ‘sensory input’ \( I \) (for example, an image). The key question within the Bayesian decoding framework is: what conditions must the stimulus-driven neural activity (\( p(r|s) \)) fulfill such that a decoder \( (p(s|r)) \) is simple, for example, linear and invariant to nuisance? For instance, imposing linearity and invariance constraints on the decoder implies constraints on tuning curves and the distribution of neural noise.

PPCs, as introduced by Ma et al.\(^{19}\), exemplify the Bayesian decoding approach. PPCs construct a Bayesian decoder that is both simple and invariant to nuisance: if a population of neurons tuned to \( s \) has ‘Poisson-like’ variability, then the optimal decoder is part of the exponential family with firing rates as natural parameters. This is a particularly convenient representation for taking products of two distributions as required by cue integration\(^{20}\) and evidence accumulation\(^{20}\). Equally important is the notion of invariance afforded by a PPC: as long as nuisance variables such as image contrast or dot coherence only multiplicatively scales tuning curves, the decoder can ignore them.

In the Bayesian decoding framework generally, imposing constraints on the decoder constrains the possible set of evoked response distributions, \( p(r|s) \). These constraints have then been formulated as predictions and tested empirically\(^{20},21\).

Contrasting Bayesian encoding and Bayesian decoding

There are four key differences between the Bayesian encoding and Bayesian decoding perspectives, addressed in the following sections: (1) what they assume the brain is inferring, (2) what the terms ‘likelihood’ and ‘posterior’ refer to, (3) the role of neural responses in the theory and (4) the empirical data and other arguments used to motivate them. As
Brain infers and encodes the same set of image features, whether viewing natural scenes or artificial stimuli in a task.\textsuperscript{6,24–26} Assuming statistical calibration in a Bayesian encoding framework also makes predictions for how the internal model should change in response to the statistics of sensory inputs during development,\textsuperscript{16} and to extensive exposure to stimuli in a particular task.\textsuperscript{17}

By contrast, Bayesian decoding models are typically applied in the context of estimating task-relevant variables. For instance, in a motion discrimination task, a Bayesian decoding question would be how the brain represents uncertainty over directions of motion. Importantly, answering this question does not require a generative model of all possible naturally occurring motions, nor does it require a true or correct reference distribution over motion directions; it requires only a statistical model of the relation between scalar motion direction in a particular task (and possibly nuisance variables such as coherence) and neural responses, \( \text{p}(\text{r}|s) \), making it observable experimentally. The difference between (typical realizations of) the Bayesian encoding and decoding perspectives is illustrated in Fig. 3.

We emphasize that Bayesian encoding typically but not necessarily involves a task-independent internal generative model, and Bayesian decoding likewise is typically but not necessarily applied to task-specific variables. In principle, we are free to apply the Bayesian decoding framework to variables in a task-independent internal model (given \( r \), what do we know about \( x \) or \( \text{p}_b(\text{X}|I) \)?), or to apply the logic of Bayesian encoding to task-specific quantities (construct \( r \) to encode a desired \( \text{p}(\text{s}|I) \)), but such examples are rare.\textsuperscript{8}

\textbf{Differing notions of likelihood.} Another difference in philosophy is evidenced by divergent usage of the term ‘likelihood’ (Fig. 3). In the typical Bayesian encoding setting, the term ‘likelihood’ is reserved for the relationship between internal model variables and sensory observations. For instance, one could speak of the ‘likelihood that this configuration of variables in the brain’s model generated the observed image,’ or \( \text{p}_b(I|x) \). This usage supports the idea that the quantity being computed is the posterior over latent variables in an internal model. In the typical Bayesian decoding setting, on the other hand, the ‘likelihood’ refers to a relationship between stimuli and neural responses, \( \text{p}(\text{r}|s) \). This usage supports the idea that the quantity of interest is the posterior over stimuli in a task. Importantly, the ‘likelihood’ in Bayesian decoding depends on choices made by the experimenter (such as their choice of stimuli), but not typically in a Bayesian encoding model.

\textbf{Differences in the relationship between distributions and neural activity.} Bayesian encoding models require two distinct assumptions: first, what is the source of the reference distribution to be encoded (for example, what is the brain’s internal model \( \text{p}_b(\text{X}, I) \)); and second, what is the linking hypothesis that maps probability distributions to neural activity (Fig. 1)? This approach of starting with the encoded distribution abstracts away from the details of neural circuits that must actually implement inference. For instance, Orban et al.\textsuperscript{14} assume that neurons in V1 encode the posterior distribution over latent variables in a Gaussian scale mixture model; this gives the reference distribution. By some unspecified mechanism, the trajectory of a set of neurons’ membrane potentials over time is assumed to trace out samples from the posterior, and that these membrane potentials elicit spikes through a nonlinear accumulation process; this gives the link to neural data. This model successfully reproduced a diverse set of known properties about V1 (ref. 24), but it is not a mechanistic model. From a modeling standpoint, the way in which an input image elicits neural activity is mediated by the reference posterior: an example of ‘downward causation.’\textsuperscript{28,29}

Bayesian decoding models, by contrast, do not distinguish between the uncertainty in an underlying probabilistic model and the uncertainty of a downstream brain area applying a decoder to some neural activity. As a result, decoding models replace the assumption of knowledge about sensory objects with knowledge about sensory variables. In this context, the notion of ‘likelihood’ is reserved for the relationship between sensory variables and neural activity (Fig. 3).

\textbf{Differences in what is assumed to be inferred.} An integral part of the Bayesian encoding framework is the existence of an abstract internal model that could in principle be implemented in silico or in the brains of other individuals or other species. Deriving predictions for neural data requires an additional linking hypothesis on the nature of distributional codes, such as whether neurons sample or encode parameters, and how either samples or parameters correspond to observable biophysical quantities such as membrane potentials and spikes. Bayesian encoding thus decomposes the question of what sensory neurons compute into two parts: first, what is the internal model that defines optimal inference (the reference distribution), and second, how do neural circuits carry out approximate inference in that model (for example, sampling or parametric)?

The brain’s internal model is typically assumed to have been calibrated through exposure to the natural world\textsuperscript{6,14} and to change only slowly with exposure to new sensory inputs in adult brains. For this reason, the generative model in Bayesian encoding models, especially in the case of early sensory areas, is often assumed to be independent of experimental context. For instance, if the brain’s internal model comprises patches of local image features, then it is assumed that the
about the link to neural activity with a constraint on the relationship between stimuli (s) and neural activity (r).

To illustrate this point, let us revisit one of the motivating examples for distributional codes of Zemel et al. and contrast the encoding and decoding approaches. Consider a rat that is placed into a water maze and must navigate to a hidden platform. Initially, the rat may be uncertain about which direction it is facing; for example, if opposite walls of the maze look the same, its correct belief about direction will be bimodal. Similarly to the orientation of a grating, head direction is a scalar variable in [0, 2π] that we will call s. In the encoding approach, one might begin by asking what is p_b(s | I) according to an internal model of the environment, where I stands for the sensory cues the rat uses to orient itself. The distribution p_b(s | I) determines how uncertain the rat ought to be, according to the internal model. One would then adopt a linking hypothesis (for example, sampling or parametric) whereby p_b(s | I) is encoded in neural activity r. In an encoding model, the encoding of a distribution may be imperfect and lossy, or it may contain more information about the distribution than is being used by downstream circuits. In either case, the way a downstream circuit uses the neural activity will generally differ from a Bayesian decoder.

Applying the Bayesian decoding framework to the same problem, we would say that the uncertainty in p(s | r) is the primary kind of uncertainty we should be concerned with, and that there is no distinction between this and the rat’s internal model. Crucially, this does not trivialize representations of uncertainty as ‘just’ a matter of optimal decoding. In the decoding approach, there may still be an ideal uncertainty that the rat ought to have when it is first placed into the maze: however, the assumption is that this uncertainty is realized through the way r is tuned to its inputs I. That is, it is left to the brain (evolution, learning) to have carefully constructed tuning functions p(r | I), such that p(s | r) is equal to p_b(s | I)^θ. One way in which the encoding and decoding perspectives can become identical, then, is when the decoded distribution p(s | r) equals the reference distribution p_b(s | I). From the decoding point of view, this requires that the encoding of p_b(s | I) into r is lossless (or efficient) in the terminology of ref. 31. From the decoding point of view, they are identical by assumption.

Finally, the preceding discussion points to an important difference between encoding and decoding philosophies in terms of how neural responses are interpreted by downstream areas. In a Bayesian decoding model, a downstream area implicitly applies Bayes’ rule to the neural responses arriving from an upstream area to extract information about a stimulus. In a Bayesian encoding model, the other hand, upstream neural activity represents samples or parameters that are then processed by the downstream area according to an underlying approximate inference algorithm, which generally will not apply Bayes’ rule to the incoming activity directly. To put it another way, if one assumes that upstream neural activity encodes samples or parameters in an approximate inference algorithm, then there is an important difference between a downstream area that interprets upstream activity as samples or as parameters (as in Bayesian encoding models), and a downstream area that decodes the activity it receives by applying Bayes’ rule to the neural activity.

Differing empirical and theoretical motivations. Finally, distinguishing Bayesian encoding and Bayesian decoding allows one to be more precise on what data and what normative arguments motivate different theories. Bayesian decoding can be motivated by the fact that humans and other species are empirically sensitive to uncertainty and prior experience, as in the classic psychophysics results on multimodal cue combination 32–36. The large literature on optimal or near-optimal Bayesian perception in controlled tasks motivates the question of how neural circuits facilitate Bayesian computations with respect to stimuli in a task, which are often scalar or low dimensional. With the additional assumption that the neural representation of task-relevant aspects of sensory inputs is formatted to be easily decoded (for example, linear and invariant to nuisance 40), this line of reasoning has given rise to predictions for neural data. These predictions have since been largely confirmed for the representation of self-motion in the dorsal medial superior temporal area 37–39. Bayesian decoding is further motivated by experimental data showing a correspondence between non-parametric likelihood functions, neural noise and behavioral indications of uncertainty 41–47.

Importantly, none of these results constitutes direct evidence for inference with respect to a (high-dimensional) internal model of natural inputs, as hypothesized in typical Bayesian encoding theories 48–51. The three lines of support for Bayesian encoding models are largely independent of the above motivations for Bayesian decoding. First, Bayesian encoding can be motivated by the purely normative argument that any rational agent that faces uncertainty ought to compute probability distributions over unobserved variables, as long as those variables directly enter into calculations of expected utility 40. Second, there is some empirical evidence that neural responses are...
self-consistent across different inputs in a manner that is predicted by having a statistically calibrated internal model. Third, there is empirical evidence for signatures of particular inference algorithms and particular internal models fit to natural inputs. This approach has been used by a series of sampling-based inference models and has successfully reproduced a wide range of neural response properties in early visual cortex. A similar approach has also been taken by parametric models, where neural circuits have been hypothesized to implement the dynamics of a variational inference algorithm. We emphasize again that existing evidence for Bayesian-like behavior in psychophysical tasks constitutes only weak evidence in support of the idea that the brain computes distributions over variables in a task-independent internal model, as usually studied in the Bayesian encoding literature.

Historically, sampling-based neural models have taken the Bayesian encoding approach, asking how neurons could sample from the posterior distribution over variables in an internal model, while PPCs—the prototypical case of parametric codes—have primarily been studied in the context of inference of low-dimensional task-relevant quantities. However, this does not reflect a fundamental distinction between the two types of distributional code. Parametric codes can and have been used in Bayesian encoding models to approximate the posterior over variables in a generative model, including PPCs, distributed distributional codes and others. Markov chain Monte Carlo sampling has been used to explain perceptual bistability, which could be seen as a form of sampling-based Bayesian decoding. To summarize, Table 1 provides a list of examples in each of the four categories defined by the sampling versus parametric and the encoding versus decoding axes. The presence of previous work in all four quadrants emphasizes that these are complementary distinctions.

Case study: primary visual cortex (V1)

We now provide a concrete example focusing on area V1. V1 has the advantage that many neurophysiological data exist, and both Bayesian encoding and Bayesian decoding approaches have enjoyed some success. We will first review existing work, and then use a simple example to show how different perspectives can lead to very different conclusions about the neural code.

Bayesian encoding models for V1

The starting point for the Bayesian encoding approach, applied to V1, is an assumption about the brain’s generative model, p(x), that is, we must specify what is x, the variable assumed to be inferred and represented by V1 neurons, and how x is related to the sensory observations I. For simple cells in area V1, Olshausen & Field proposed a linear Gaussian likelihood I ~ N(Ax, δ2) with a sparse independent prior p(x) as the brain’s internal model. (Here, I ~ N(μ, Σ) indicates that I is a random variable drawn from a multivariate normal distribution, and μ (μ, Σ) denotes its density function.) The observed retinal image, I, is assumed to be a linear combination of ‘projective fields’ (PF) plus unexplained pixel noise Σ. The matrix A is a feature dictionary with projective fields as its columns: A = (PF1, ..., PFk). Each of the n projective fields is weighted by the corresponding element of x = (x1, ..., xn). Intuitively, in this model, V1 activity is assumed to represent beliefs about what values for x explain a given retinal image, I.

The next assumption in the encoding framework is about how the posterior distribution, p(x|I), is represented by neural responses, r. Olshausen & Field assumed that each x was represented by a single neuron whose firing rate was proportional to the most probable value for x, given an image (maximum a posteriori, MAP): r ~ argmax x p(x|I). In this model, a single neuron represents the most likely intensity with which a visual feature is present in the image. This is not a fully Bayesian encoding model in the sense that only the MAP, but not the full posterior distribution p(x|I), is encoded in neural responses. Empirical support for this model is based on the observation that learning (fitting) this model on natural images yields visual features (PFs) that are localized, oriented and band-pass filtered, implying neural responses and receptive fields with similar properties—just as observed empirically.

Subsequent work has both modified and extended this generative model and combined it with different neural codes. Hoyer & Hyvärinen proposed that neural responses can be understood as samples from the posterior in the same generative model to qualitatively explain the variability and mean–variance relationship of neural responses. Schwartz & Simoncelli extended the generative model to a Gaussian scale mixture model to explain the empirically observed contrast normalisation of V1 responses, and Orban et al. found agreement between the predictions of a Gaussian scale mixture model combined with neural sampling and a wide range of observations related to the stimulus dependence of neural variability. Bornschein et al. proposed a variation of the generative model of Olshausen & Field using a nonlinear Gaussian likelihood and/or binary as opposed to continuous latents. Coen-Cagli et al. found that a further extension to the generative model in the form of a mixture of Gaussian scale mixture model could explain center–surround interactions in V1, and Banyai et al. explained noise correlations by higher-order image structure. Finally, Hafner et al. combined the generative model of Olshausen & Field with the ideal observer model of a discrimination task to explain choice probabilities and task-dependent noise correlations of V1 neurons.

The key shared element of all these models is an explicit assumption about the computational variable x that is being represented, and how this variable is related to the sensory observations I. This model being adapted to natural inputs is an important constraint, and the model parameters are usually obtained by fitting the model to sets of natural images. These models are then image-computable and can be queried using natural inputs or images presented in a task.

Bayesian decoding models for V1

The starting point for the Bayesian decoding approach, applied to V1, is a measurement of the conditional probability, p(r|I), for some stimulus s to which V1 neurons are tuned, and that is hypothesized to be represented, such as orientation. Importantly, this means that the measured likelihood is to some extent under experimental control, because the experimenter chooses what images correspond to each value of s (for example, the size, contrast or spatial frequency of a grating). In general, for an arbitrary s, this likelihood

---

Table 1: Classification of existing models

| Perspective | Bayesian encoding | Bayesian decoding |
|-------------|-----------------|------------------|
| Sampling-based representation | Hoyer & Hyvärinen, Pecovsky et al., Berkes et al., Buesing et al., Gershman et al., Savin & Denève, Probst et al., Orban et al., Haefner et al., Aitchison & Lengyel, Festa et al., Banyai et al., Echeveste et al. | Moreno-Bote et al. |
| Parametric representation | Zemel et al., Sahani & Dayan, Friston, George et al., Beck et al., Raju et al., Vertes et al., Tajima et al. | Ma et al., Beck et al., Beck et al., Hou et al., Tajima et al., Moreno-Bote et al. |

Classifying previous work on Bayesian neural models according to whether they construct Bayesian encoding or decoding models, and whether they use a sampling-based or a parametric neural representation. Tajima et al. contains elements of both encoding and decoding. Moreno-Bote et al. contains elements of both sampling-based and parametric decoding.
will be very complicated, reflecting the fact that \( s \) cannot easily be decoded from \( r \) (for example, object identity from V1). However, for V1 responses it has empirically been found that the optimal Bayesian decoder for orientation is approximately linear in spike counts and invariant to contrast, a classic nuisance variable\(^\text{15}\). This finding has been interpreted as meaning that V1 activity ‘represents’ orientation. In conjunction with the Poisson-like neural response variability in V1, this implies that the beliefs of a Bayesian decoder of orientation applied to the neural responses are part of the exponential family. This makes V1 a PPC over orientation\(^\text{19}\).

The same logic applies to other candidates for \( s \) that modulate the responses of V1 neurons, such as spatial frequency or location. The key element of the Bayesian decoding approach is taking the perspective of downstream circuits trying to extract information about \( s \) from V1 activity: how is the information about \( s \) formatted in V1 activity, and is \( p(s|r) \) ‘simple’? In contrast to the Bayesian encoding perspective, which justifies its choice of \( x \) by its fit to natural images and its ability to predict neural responses, the Bayesian decoding perspective justifies its choice of \( s \) by desirable properties of an efficient decoder, for example, linearity and invariance to nuisance variables.

**Example model where decoding a stimulus \( s \) from encoded samples results in a PPC.** As our main points are conceptual in nature, we will develop the link between the encoding and the decoding approach for the simple case of a linear Gaussian model with a Gaussian prior, under the assumption of a sampling-based neural code. These simplifying assumptions make the difference between Bayesian encoding and Bayesian decoding clear and analytically tractable, but are not meant to maximize biological plausibility. The core of our argument is that a Bayesian encoding model based on one type of neural code (for example, sampling) and a Bayesian decoding model based on another type (for example, parametric) are not necessarily contradictory, and instead offer complementary perspectives on the same system.

Given an image \( I \), we assume that V1 neurons encode the posterior \( p_b(x|I) \) by sampling \( r \) values from the posterior distribution, \( x^{(i)} \sim p_b(x|I) \propto p_b(I|x)p_b(x) \) where \( p_b(x) \) is the brain’s prior over \( x \). We assume that responses from a population of \( n \) neurons correspond to samples from the posterior over \( x \), so that at each instant, the population response, \( r^{(i)} \), equals the sample \( x^{(i)} \). Each sample of \( (x_i, r_i) \) represents the brain’s instantaneous belief about the intensity of the feature PF in the image.

We will now apply the Bayesian decoding approach to the sequence of samples produced by the sampling-based encoding model described above. An ideal observer applies Bayes’ rule to infer \( p(s|r^{(1)}, \ldots, r^{(T)}) \) using knowledge of the probabilistic relationship between samples at time \( t \) (\( x^{(t)} \) or \( r^{(t)} \)) and \( s \):

\[
p(s|r^{(1)}, \ldots, r^{(T)}) \propto p(s)p_b(r^{(1)}, \ldots, r^{(t)}|s) \propto s p(s) \int p(I|s)p_b(r^{(1)}, \ldots, r^{(t)}|I)dI. \quad (1)
\]

That is, the optimal decoder combines knowledge of (i) how likely an image \( I \) is to generate a set of samples of \( x \) (or \( r \)), and (ii) how likely a stimulus value \( s \) is to generate an image \( I \). In general, this decoded distribution over \( s \) may be arbitrarily complex and intractable. One factor that is under experimental control is the ‘template’ function \( T(s) \) which renders an image, such as a grating with orientation \( s \). This provides the link between \( s \) and \( I \) in equation (1). In our model, we assume that the input the brain receives is a noisy version of that template (Fig. 4).

The first simplification to the general form of the optimal decoder in equation (2) we can derive, under the assumption of a Gaussian likelihood, is that the posterior over \( s \) depends only on the mean rate of \( r \) (that is, a rate code rather than temporal code):
where \( \bar{r} = \frac{1}{N_r} \sum_{i=1}^{N_r} r(i) \) is the mean response after \( r \) samples (Supplementary Note). Any decoder that obeys equation (2) can be seen as a kind of parametric code over \( s \), where the rates \( r \) are the parameters. A second convenient property for a decoder to have is if the optimal decoder is in an exponential family, or

\[
p(s|\bar{r}) = p(s|\bar{r})
\]

where \( f(r) = \frac{1}{N_r} \sum_{i=1}^{N_r} r(i) \) is the mean response after \( r \) samples (Supplementary Note). Any decoder that obeys equation (2) can be seen as a kind of parametric code over \( s \), where the rates \( r \) are the parameters. A second convenient property for a decoder to have is if the optimal decoder is in an exponential family, or

\[
p(s|\bar{r}) = p(s|\bar{r}) = \exp \left( \sum_{i=1}^{N_r} r(i) \right)
\]

Whenever equation (3) is true, then we would say that the neural activity forms a particular kind of parametric code called a nonlinear PPC over \( s \). A final convenient property for a decoder to have is if \( f(r) \) is linear:

\[
p(s|\bar{r}) = p(s|\bar{r}) \propto g(s) \exp \left( \sum_{i=1}^{N_r} r(i) \right)
\]

This is the definition of a linear PPC over \( s \), but note that PPCs also place restrictions on nuisance variables, which we have omitted here.

In our simplified encoding model, we indeed find that the optimal decoder is, in fact, a linear PPC over \( s \) as defined in equation (4) (derivation in the Supplementary Note). This sequence of steps from equation (1) through (4) suggests a general way to derive the Bayesian decoding model implied by a given Bayesian encoding model. See Shikumara et al. 1 for further discussion of the relation between this PPC and the brain’s internal model.

**Discussion**

We have identified a previously unstated difference between approaches to constructing Bayesian neural models: Bayesian encoding and Bayesian decoding. This distinction provides new insights into the long-standing debate about the nature of the neural code. Importantly, we have demonstrated that these two approaches can give rise to different but compatible models of the same neural circuit. The complementary nature of these approaches has direct implications for both theoretical debates and the correct interpretation of empirical data, for example, the much-debated question of whether neural responses are more closely related to parameters of the encoded probability distribution, as in PPCs 9 and in distributed distributional codes 47, or to samples from the distribution, as in neural sampling (reviewed in refs. 4, 5, 17, 18).

Our model is constructive proof that Bayesian encoding and Bayesian decoding models can be compatible on the same data, but this will not be true in general. For instance, non-Gaussian encoding models will not generally form a linear PPC from the decoding perspective, or only for specific sets of stimuli, or they may be decodable only as a nonlinear PPC. Generalizing from our specific example, the key question is which families of Bayesian encoding models, consisting of both \( p_s(\mathbf{x}|s) \) and an assumption about the link to neural responses, are compatible with which families of Bayesian decoding models, consisting of \( p_r(s|r) \) and \( p(r) \). These will come in pairs—each family of Bayesian encoding models defines a family of compatible Bayesian decoding models, and vice versa. Identifying these pairs of compatible model families is an open theoretical question with important implications for the interpretation of empirical data.

Our arguments also raise questions about what makes a neural code ‘distributional’, that is, representing a whole distribution rather than just a point estimate, and what would constitute empirical evidence for it. Consider a variant of our example model in which neural responses are proportional to a point estimate of \( x \) such as its mean or mode rather than samples from the posterior. This would be a poor Bayesian encoding model in the sense that the full \( p_s(\mathbf{x}|s) \) distribution is not recoverable from \( r \). Yet even this reduced model gives rise to a distributional code over \( s \) (a PPC, in fact). Such a point estimate code over variables in the brain’s internal model would still enable many of the apparently Bayesian behaviors observed in low-dimensional psychophysics tasks and used to motivate Bayesian decoding theories, as discussed in above (see also ref. 58). It would therefore be a mistake to treat empirical evidence for Bayesian behavior in a particular task alone as evidence that the brain represents probability distributions over variables in an internal model of sensory inputs 48,59. The distinction between Bayesian encoding and Bayesian decoding might thus productively add to the open philosophical question: ‘If behavior is probabilistic, why does perception not seem probabilistic?’ 59,60.

The seminal paper by Zemel et al. 9 introduced the concept of encoding (and decoding) general probability distributions in (and from) neural activity. Most work over the following 20+ years typically focused on either encoding or decoding (Table 1), despite Zemel et al. 9 considering both perspectives as tightly linked. This divergence was likely strengthened because encoding studies almost exclusively considered internal latent variables \( x \), while work taking the decoding perspective considered distributions over task-defined variables \( s \). From today’s perspective, the encoding formalism of Zemel et al. 9 and its application in Sahani & Dayan 41 maps naturally onto our Bayesian encoding category. Furthermore, it is philosophically closely aligned with the other studies in this category, and shares with them the idea that implied decoders that are non-Bayesian. Interestingly, while Zemel et al. 9 discounted the possibility of optimally decoding the encoded distribution using Bayes’ rule as too inflexible, almost all later studies that took the decoding approach were based on Bayes’ rule, and now form our Bayesian decoding category.

The key step in our example system above that allowed us to interpret samples of \( x \) as a PPC was to construct the PPC over a different variable: \( s \). This raises the question: what if \( s \) is part of the brain’s internal model? One possibility is that ‘orientation’ (or any other \( s \) in a task) is a useful abstraction of natural inputs, in which case it may have been learned (or evolved) and may permanently be a part of the brain’s internal model. Another possibility is that orientation (or any other \( s \)) is part of the brain’s internal model because the brain changes its internal model as the result of learning the present task. 9,57,22 Echoing that mentioned above, even if \( s \) is part of the brain’s internal model, Bayesian encoding and Bayesian decoding models would nonetheless differ in their approach to the question of how neural responses, \( r \), relate to the distribution on \( s \). Bayesian encoding models would begin with a generative model of sensory input \( \mathbf{I} \) from \( s \) (and possible other internal variables \( x \)) and ask how the true posterior \( p_r(s|\mathbf{I}) \) is represented by neural responses \( r \). Bayesian decoding models, on the other hand, would investigate the relationship between \( s \) in the world and evoked neural responses, \( p(r|s) \), and study a different kind of posterior, \( p(s|r) \), which takes the perspective of the experimenter, or possibly the rest of the brain trying to read out \( s \) from \( r \). If the decoded distribution, \( p(s|r) \), matches the ideal encoded distribution, \( p_r(s) \), then the code for \( s \) is said to be efficient 91.

The choice of variable, which is assumed to be inferred, also affects the interpretation of neural variability. In our example above, neural variability is directly related to the uncertainty in the posterior over \( x \). In contrast, the uncertainty over \( s \) encoded by the Bayesian decoding model is unrelated to the neural variability, depending on the samples only through their mean, rather than their variance. This is an important point for experiments that seek to test the neural sampling hypothesis by relating neural variability and ‘uncertainty’ in our example model, uncertainty only over \( x \) but not over \( s \) manifests as neural variability, while \( s \) is the variable most commonly and naturally manipulated in an experiment.

The issues raised in this paper for models of visual perception also have implications for Bayesian models of cognition, where ideas related to sampling 28-30, variational inference 33,35,59 or both 40 have been invoked to explain a wide variety of heuristics and biases (reviewed in refs. 17,72). Here, too, it is important to distinguish between probabilistic models of the world that are posited to exist in a person’s mind (as is typical in
Bayesian encoding) from experimenter-defined models of a particular task (as is typical in Bayesian decoding). Closely related is the distinction drawn by Knill & Richards between the ‘inference problem’ (what the brain infers in the internal model it assumes) and the ‘information problem’ (what information is available in the world). For example, Vul et al. argue that certain deviations from Bayes-optimal behavior can be explained as the result of basing decisions on a single Monte Carlo sample. However, it is conceivable that what appears to be a single point-estimate sample over a quantity relevant to a task may, in fact, be a local, perhaps unimodal distribution over a detailed internal model, as in parametric approximations. It is further conceivable that multiple ‘samples’ correspond to a mixture of parametric approximations over an internal model. Conversely, a single high-dimensional point estimate of an internal model may be sufficient to facilitate apparently Bayesian behavior with respect to a low-dimensional task. Changing our reference frame from internal models to experimenter-defined tasks may make samples appear as parameters, or vice versa.

In complementary work, Koblinger et al. recently posed the related but different question of whether uncertainty in the brain is represented ‘constitutively’ about many variables regardless of their relevance for a specific task, or ‘opportunistically’ about task-relevant variables. While encoding and decoding models have primarily been applied in task-independent and task-dependent contexts, respectively, one can imagine both ‘constitutive’ and ‘opportunistic’ encoding and decoding models. Walker et al. pointed out a distinction between ‘correlational’ and ‘code-driven’ methodologies to the study of neural representations of uncertainty.

**Code availability**

Two panels in Fig. 4 were generated by simulation. The code is available at https://github.com/haefnerlab/bayesian-encoding-decoding/.

**References**

1. von Helmholtz, H. Treatise on Physiological Optics (The Optical Society of America, 1925).
2. Knill, D. C. & Richards, W. Perception as Bayesian Inference (Cambridge Univ. Press, 1996).
3. Kersten, D., Mamassian, P. & Yuille, A. Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).
4. Fiser, J., Berkes, P., Orbán, G. & Lengyel, M. Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn. Sci.* **14**, 119–130 (2010).
5. Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16**, 1170–1178 (2013).
6. Shikvumar, S., Lange, R. D., Chattoraj, A. & Haefner, R. M. A probabilistic population code based on neural samples. *Adv. Neural Inf. Process. Syst.* **31**, 7070–7079 (2018).
7. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
8. Olshausen, B. A. & Field, D. J. Sparse coding with an incomplete basis set: a strategy employed by V1? *Vis. Res.* **37**, 3311–3325 (1997).
9. Zemel, R. S., Dayan, P. & Pouget, A. Probabilistic interpretation of neural codes. *Neural Comput.* **10**, 403–430 (1998).
10. Berkes, P., Orbán, G., Lengyel, M. & Fiser, J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* **331**, 83–87 (2011).
11. Dayan, P., Hinton, G. E., Neal, R. & Zemel, R. The Helmholtz machine. *Neural Comput.* **7**, 889–904 (1995).
12. Dayan, P. & Abbott, L. F. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (MIT Press, 2001).
13. Bell, A. J. & Sejnowski, T. J. The ‘independent components’ of scenes are edge filters. *Vis. Res.* **37**, 3327–3338 (1997).
14. Murphy, K. P. *Machine Learning: a Probabilistic Perspective* (MIT Press, 2012).
15. Wainwright, M. J. & Jordan, M. I. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1**, 1–305 (2008).
16. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
17. Sanborn, A. N. Types of approximation for probabilistic cognition: sampling and variational. *Brain Cogn.* **112**, 98–101 (2015).
18. Gershman, S. J. & Beck, J. M. in *Computational models of brain and behavior* (ed. A. Moustafa), 453–466 (Wiley-Blackwell, 2016).
19. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432–1438 (2006).
20. Beck, J. M. et al. Probabilistic population codes for Bayesian decision-making. *Neuron* **36**, 1142–1152 (2008).
21. Fetsch, C. R., Pouget, A., DeAngelis, G. C. & Angelaki, D. E. Neural correlates of reliability-based cue weighting during multisensory integration. *Nat. Neurosci.* **15**, 146–154 (2011).
22. Fetsch, C. R., DeAngelis, G. C. & Angelaki, D. E. Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. *Nat. Rev. Neurosci.* **14**, 429–442 (2013).
23. Hou, H., Zheng, Q., Zhao, Y., Pouget, A. & Gu, Y. Neural correlates of optimal multisensory decision making under time-varying reliabilities with an invariant linear probabilistic population code. *Neuron* **104**, 1010–1021 (2019).
24. Orbán, G., Berkes, P., Fiser, J. & Lengyel, M. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron* **92**, 530–543 (2016).
25. Haefner, R. M., Berkes, P. & Fiser, J. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron* **90**, 649–660 (2016).
26. Bényai, M. et al. Stimulus complexity shapes response correlations in primary visual cortex. *Proc. Natl Acad. Sci. USA* **116**, 2723–2732 (2019).
27. Lange, R. D. & Haefner, R. M. Task-induced neural covariability as a signature of approximate Bayesian learning and inference. *PLoS Comput. Biol.* **18**, e1009557 (2022).
28. Campbell, D. T. in *Studies in the Philosophy of Biology* (ed. F. J. Ayala) 179–186 (Macmillan, 1974).
29. Yabio, S. Mental causation. *Philos. Rev.* **101**, 245–280 (1992).
30. Morris, R. Developments of a water-maze procedure for studying spatial learning in the rat. *J. Neurosci. Methods* **11**, 47–60 (1984).
31. Beck, J. M., Keller, H. & Pouget, A. Complex inference in neural circuits with probabilistic population codes and topic models. *Adv. Neural Inf. Process. Syst.* **25**, 3068–3076 (2012).
32. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
33. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
34. Alais, D. & Burr, D. The ventroloquistic effect results from near-optimal bimodal integration. *Curr. Biol.* **14**, 257–262 (2004).
35. Körding, K. P. Decision theory: what ‘should’ the nervous system do? *Science* **318**, 606–610 (2007).
36. Angelaki, D. E., Gu, Y. & DeAngelis, G. C. Multisensory integration: psychophysics, neurophysiology, and computation. *Curr. Opin. Neurobiol.* **19**, 452–458 (2009).
37. Walker, E. Y., Cotton, R. J., Ma, W. J. & Tolias, A. S. A neural basis of probabilistic computation in visual cortex. *Nat. Neurosci.* **23**, 122–129 (2019).
38. Rahnev, D. The Bayesian brain: what is it and do humans have it? *Behav. Brain Sci.* **42**, e238 (2019).
39. Koblinger, Á., Fiser, J. & Lengyel, M. Representations of uncertainty: where art thou? Curr. Opin. Behav. Sci. 38, 150–162 (2021).
40. Jaynes, E. T. Probability Theory: the Logic of Science (Cambridge Univ. Press, 2003).
41. Aitchison, L. & Lengyel, M. The hamiltonian brain: efficient probabilistic inference with excitatory–inhibitory neural circuit dynamics. PLoS Comput. Biol. 12, e1005186 (2016).
42. Echeveste, R., Aitchison, L., Hennequin, G. & Lengyel, M. Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. Nat. Neurosci. 23, 1138–1149 (2020).
43. George, D. & Hawkins, J. Towards a mathematical theory of cortical micro-circuits. PLoS Comput. Biol. 5, e1000532 (2009).
44. Raju, R. V. & Pitkow, X. in Advances in Neural Information Processing Systems vol. 30 (eds. Lee, D. D. et al.) 2029–2037 (Curran Associates, 2016).
45. Grabska-Barwinska, A., Beck, J. M., Pouget, A. & Latham, P. E. in Advances in Neural Information Processing Systems vol. 26 (eds. Burges, C. J. C. et al.) 1968–1976 (Curran Associates, 2013).
46. George, D. et al. Cortical microcircuits from a generative vision perspective. Nat. Neurosci. 26, 883–893 (2016).
47. Friston, K. J. A theory of cortical responses. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360, 815–836 (2005).
48. Lavin, A., Guntpalli, J. S., Lázaro-gredilla, M., Lehrach, W. & George, D. Explaining visual cortex phenomena using recursive cortical network. Preprint at bioRxiv https://doi.org/10.1101/380048 (2018).
49. Vertes, E. & Sahani, M. in Advances in Neural Information Processing Systems vol. 31 (eds Bengio, S. et al.) (Curran Associates, 2018).
50. Moreno-Bote, R., Knill, D. C. & Pouget, A. Bayesian sampling in cortical microns. PLoS Comput. Biol. 9, e1003062 (2013).
51. Gershman, S. J., Vul, E. & Tenenbaum, J. B. The anchoring bias reflects rational use of cognitive resources. Psychon. Bull. Rev. 25, 322–334 (2017).
52. Hohwy, J., Roepstorff, A. & Friston, K. J. Predictive coding explains binocular rivalry: an epistemological review. Cogn. Sci. 38, 677–701 (2008).
53. Hoyer, P. O. & Hyvärinen, A. Interpreting neural response variability as monte carlo sampling of the posterior. Adv. Neural Inf. Process. Syst. 17, 293–300 (2003).
54. Schwartz, O. & Simoncelli, E. P. Natural signal statistics and sensory gain control. Nat. Neurosci. 4, 819–825 (2001).
55. Bornchein, J., Henniges, M. & Lücke, J. Are V1 simple cells optimized for visual occlusions? a comparative study. PLoS Comput. Biol. 9, e1003062 (2013).
56. Coen-Cagli, R., Kohn, A. & Schwartz, O. Flexible gating of contextual influences in natural vision. Nat. Neurosci. 18, 1648–1655 (2015).
57. Graf, A. B. A., Kohn, A., Jazayeri, M. & Movshon, J. A. Decoding the activity of neuronal populations in macaque primary visual cortex. Nat. Neurosci. 14, 239–247 (2011).
58. Orhan, A. E. & Ma, W. J. Efficient probabilistic inference in generic neural networks trained with non-probabilistic feedback. Nat. Commun. 8, 138 (2017).
59. Block, N. If perception is probabilistic, why does it not seem probabilistic? Philos. Trans. R. Soc. Lond. B Biol. Sci. 373, 20170341 (2018).
60. Rahnev, D., Block, N., Jehee, J. & Denison, R. Is perception probabilistic? Clarifying the definitions. Preprint at https://psyarxiv.com/f8v5r/ (2020).
61. Sahani, M. & Dayan, P. Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. Neural Comput. 15, 2255–2279 (2003).
62. Vul, E. & Rich, A. N. Independent sampling of features enables conscious perception of bound objects. Psychol. Sci. 21, 1168–1175 (2010).
63. Sanborn, A. N., Griffiths, T. L. & Navarro, D. J. Rational approximations to rational models: alternative algorithms for category learning. Psychol. Rev. 117, 1144–1167 (2010).
64. Lieder, F., Hsu, M. & Griffiths, T. L. The high availability of extreme events serves resource-rational decision-making. In Proc. 36th Annual Conference of the Cognitive Science Society (eds. Bello, P. et al.) 2567–2572 (Cognitive Science Soc., 2014).
65. Vul, E., Goodman, N. D., Griffiths, T. L. & Tenenbaum, J. B. One and done? Optimal decisions from very few samples. Cogn. Sci. 38, 599–637 (2014).
66. Sanborn, A. N. & Chater, N. Bayesian brains without probabilities. Trends Cogn. Sci. 20, 893–893 (2016).
67. Lieder, F., Griffiths, T. L., Huys, Q. J. M. & Goodman, N. D. The anchoring bias reflects rational use of cognitive resources. Psychon. Bull. Rev. 25, 322–334 (2017).
68. Zhu, J.-Q., Sanborn, A. N. & Chater, N. The Bayesian sampler: generic Bayesian inference causes incoherence in human probability judgments. Psychol. Rev. 127, 719–748 (2020).
69. Daw, N. D., Courville, A. C. & Dayan, P. in The Probabilistic Mind: Prospects for Bayesian Cognitive Science (eds. Chater, N. & Oaksford, M.) https://doi.org/10.1093/acprof:os o/9780199216093.003.0019 (Oxford Scholarship Online, 2008).
70. Sanborn, A. N. & Silva, R. Constraining bridges between levels of analysis: a computational justification for locally Bayesian learning. J. Math. Psychol. 57, 94–106 (2013).
71. Lange, R. D., Chattoraj, A., Beck, J. M., Yates, J. L. & Haefner, R. M. A confirmation bias in perceptual decision-making due to hierarchical approximate inference. PLoS Comput. Biol. 17, e1009517 (2021).
72. Griffiths, T. L., Vul, E. & Sanborn, A. N. Bridging levels of analysis for probabilistic models of cognition. Curr. Dir. Psychol. Sci. 21, 263–268 (2012).
73. Jaakkola, T. S. & Jordan, M. I. in Learning in Graphical Models (ed. Jordan, M. I.) 163–174 (Kluwer, 1998).
74. Lange, R. D., Benjamin, A. S., Haefner, R. M. & Pitkow, X. Interpolating between sampling and variational inference with infinite stochastic mixtures. Proc. Machine Learn. Res. 180, 1063–1073 (2022).
75. W"{a}lker, E. Y. et al. Studying the neural representations of uncertainty. Nat. Neurosci. 26, 1857–1867 (2023).
76. Pecevski, D., Buesing, L. & Maass, W. Probabilistic inferences general graphical models through sampling in stochastic networks of spiking neurons. PLoS Comput. Biol. 7, e1002294 (2011).
77. Buesing, L., Bill, J., Nessler, B. & Maass, W. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. PLoS Comput. Biol. 7, e1002211 (2011).
78. Savin, C. & Denève, S. Spatio-temporal representations of uncertainty in spiking neural networks. In Advances in Neural Information Processing Systems vol. 27 2024–2032 (NIPS, 2014).
79. Probst, D. et al. Probabilistic inference in discrete spaces can be implemented into networks of LIF neurons. Front. Comput. Neurosci. 9, 13 (2015).
80. Festa, D., Aschner, A., Davila, A., Kohn, A. & Coen-Cagli, R. Neuronal variability reflects probabilistic inference tuned to natural image statistics. Nat. Commun. 12, 3635 (2021).
81. Tajima, C. I. et al. Population code dynamics in categorical perception. Sci. Rep. 6, 22536 (2016).
82. Beck, J. M., Latham, P. E. & Pouget, A. Marginalization in neural circuits with divisive normalization. J. Neurosci. 31, 15310–15319 (2011).
Acknowledgements
We thank the many colleagues with whom we have discussed the ideas in this paper, especially M. Lengyel and J. Drugowitsch for their detailed comments on an earlier version of this manuscript. This work was supported by the National Institutes of Health (NIH) R01 grant EY028811, NIH U19 grant 1U19NS118246-01 and a National Science Foundation CAREER grant IIS-2143440 to R.M.H.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-023-01458-6.

Correspondence should be addressed to Richard D. Lange.