nature neuroscience

Review article

https://doi.org/10.1038/s41593-023-01444-y

Studying the neural representations of uncertainty

Received: 8 February 2022

Accepted: 30 August 2023

Published online: 9 October 2023

Check for updates

Edgar Y. Walker[1,9], Stephan Pohl[2,9], Rachel N. Denison[3], David L. Barack[4,5], Jennifer Lee[6], Ned Block[2], Wei Ji Ma[6,7,10] & Florent Meyniel[8,10] ✉

The study of the brain's representations of uncertainty is a central topic in neuroscience. Unlike most quantities of which the neural representation is studied, uncertainty is a property of an observer's beliefs about the world, which poses specific methodological challenges. We analyze how the literature on the neural representations of uncertainty addresses those challenges and distinguish between 'code-driven' and 'correlational' approaches. Code-driven approaches make assumptions about the neural code for representing world states and the associated uncertainty. By contrast, correlational approaches search for relationships between uncertainty and neural activity without constraints on the neural representation of the world state that this uncertainty accompanies. To compare these two approaches, we apply several criteria for neural representations: sensitivity, specificity, invariance and functionality. Our analysis reveals that the two approaches lead to different but complementary findings, shaping new research questions and guiding future experiments.

Understanding how the brain represents its environment is one of the major goals of neuroscience and psychology. Another major goal is to understand the uncertainty of these representations[1–5]. Taking into account uncertainty in perceptual processing can be crucial when interacting with the world. Imagine that while hiking through the mountains you have to decide whether to attempt to cross a steep slope. Besides your perception of the slope itself, your uncertainty about the slope should also be taken into account. Perhaps you should move closer in order to reduce your uncertainty before you decide to attempt the crossing. A wide range of human behavior takes into account such uncertainty, including decision-making[6–9], learning[10–13], perception[3,14–16] including multisensory fusion[17–20], motor control[21,22] and memory[23–26]. Similar observations have been made in nonhuman animals[27–35].

Many neuroscientists aim to understand how this uncertainty is represented in the brain. Studies of uncertainty often contain claims of the form 'in a given brain region, neural activity r represents uncertainty about the latent state s'. In practice, r can be measured with functional magnetic resonance imaging (fMRI), electroencephalography, intracranial recordings of local field potentials or spike trains, among others, and s can be the orientation of an object, a reward probability or some other feature. The goal of this Review is to provide a framework for categorizing and evaluating claims about the representation of uncertainty.

Defining uncertainty

Uncertainty characterizes the representation of a world state by an observer

Consider some subject who perceives s, some feature of interest of the world state. We will understand this situation in terms of a generative model (Fig. 1a; see glossary in the Supplementary Information). The feature s is not directly accessible, and is thus called a 'latent' feature.

[1]Department of Physiology and Biophysics, Computational Neuroscience Center, University of Washington, Seattle, WA, USA. [2]Department of Philosophy, New York University, New York, NY, USA. [3]Department of Psychological & Brain Sciences, Boston University, Boston, MA, USA. [4]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, USA. [5]Department of Philosophy, University of Pennsylvania, Philadelphia, PA, USA. [6]Center for Neural Science, New York University, New York, NY, USA. [7]Department of Psychology, New York University, New York, NY, USA. [8]Cognitive Neuroimaging Unit, INSERM, CEA, CNRS, Université Paris-Saclay, NeuroSpin center, Gif-sur-Yvette, France. [9]These authors contributed equally: Edgar Y Walker, Stephan Pohl. [10]These authors jointly supervised this work: Wei Ji Ma, Florent Meyniel. ✉e-mail: florent.meyniel@cea.fr

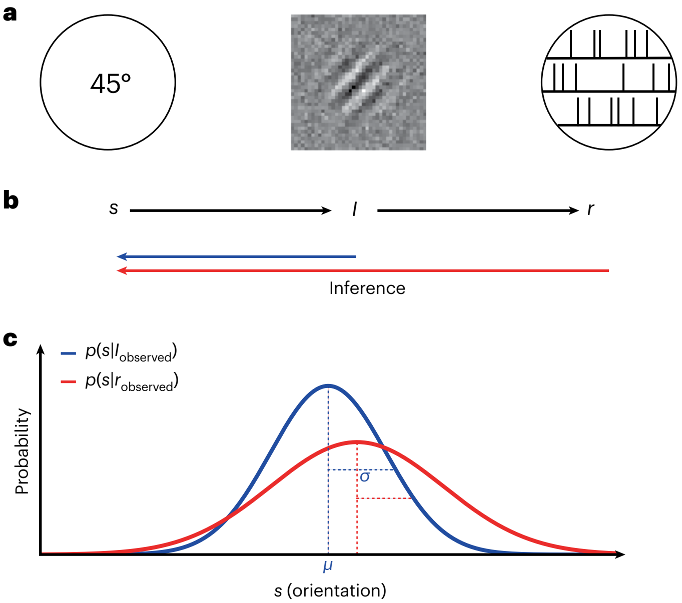

Fig. 1 | Uncertainty from a generative model. a, Example generative model. The world state s is an orientation. The input I is an image of a grating with orientation s and some pixel noise. The neural response r is a train of spikes from neurons in some population. **b**, Black arrows indicate the dependencies in the generative model between the world state s, input I and neural response r. Red and blue arrows indicate inferences an observer could make by inverting dependencies in the generative model from I and r, respectively. **c**, Probability distributions of s given the particular input I_{observed} or the particular neural response r_{observed} in a given trial. σ is a measure of the uncertainty about s given I_{observed}, and μ is the expected value of s given I_{observed}.

BOX 1

Measures of uncertainty

Different formal measures are available to summarize uncertainty u. Often, uncertainty is understood as the standard deviation of a random variable under a posterior distribution, especially for the frequently used Gaussian distribution (Fig. 1c). The larger the standard deviation (that is, the more spread out the probability distribution) the more uncertainty there is about this variable.

Another useful measure of uncertainty is entropy. The entropy[37] of the posterior distribution $p(s|I_{\text{observed}})$ is a measure for how much freedom of choice (hence, uncertainty) there is left about the variable s, after one already knows that the variable I takes the value I_{observed}. The advantage of entropy as a measure of uncertainty is that it applies to probability distributions of any shape (categorical and numeric variables, and with one or more dimensions). However, the fact that entropy ignores ordinality can be a disadvantage: for instance, if two orientations have high probability and all other orientations have the same low probability, entropy (unlike standard deviation) will be the same no matter whether those two orientations are very close or very far apart.

Yet rather than summarizing uncertainty in a single quantity, one might also keep track of it implicitly in terms of the full probability distribution. If one were to represent the state of the world s in terms of, for instance, the posterior distribution $p(s|I_{\text{observed}})$, uncertainty about s would be implicit in that representation. This uncertainty can be taken into account implicitly by performing computations on the full probability distribution[92]. Whether observers use full distributions or summaries is an open empirical question[58,108,124].

The observer receives information about s from the more proximal input state I. The brain processes this input to arrive at the neural response r, which is a representation of s.

In an experimental context, I is the input to the observer that is generated by the feature s in a particular trial. For example, in a visual task, I is the pattern of light that hits the observer's retina, which in practice is considered to be equivalent to the pattern of pixels presented on a screen. As a standard example throughout, let I be an image: a grating with orientation s and added random pixel noise (Fig. 1a). Participants report their estimate of the orientation s.

Consider an observer who forms a representation of the world state s through a process that can be described as an inference from I (Fig. 1b). That is, the observer computes values for s given the observed I (I_{observed}) and the dependence of I on s assumed in the generative model[36]. However, one and the same input can often be generated from multiple states of the world. That is, the state of the world is under-determined given the input, leaving the observer uncertain about s.

Unlike most quantities that are represented in perception, the uncertainty u about s is not a world state. Rather, the uncertainty u is a property of an observer's belief about the world; u measures the lack of information the observer has about s on the basis of an inference from I[37].

The posterior probability distribution $p(s|I_{\text{observed}})$ characterizes the uncertainty about s given I_{observed} in a given trial (Fig. 1c). This distribution describes the probability of different values of s given the particular input I_{observed} (Box 1). If there were no uncertainty, the value of s would be perfectly determined by some particular input I_{observed}, and all the probability mass in the distribution $p(s|I_{\text{observed}})$ would be assigned to a single value of s. But here, multiple values of s are possible given I_{observed}; thus, there is uncertainty about s given I_{observed}.

The generative model assumed in an observer's inference need not be the true generative model. The generative process might be too complex for observers to compute an optimal inference[38,39]; they might, for instance, exclude some variables and simplify the shapes of probability functions[40–42]. Because an observer's uncertainty depends on the generative model assumed in their inference, a major task in the study of representations of uncertainty is to study what generative models are assumed by observers.

Often, an idealized observer is considered who infers s from I based on an optimal inference and the true generative model; the uncertainty such an idealized observer would have about s is called the ideal observer uncertainty[5,14,43].

Origins of uncertainty

Uncertainty about the world state s is present whenever s is under-determined given the inference an observer performs. This under-determination manifests as a many-to-one mapping from the world state to a later state in the generative model assumed by the observer.

One source of under-determination is ambiguity of the input, illustrated for instance by the case of the Necker cube[44]. One and the same two-dimensional image could be interpreted to be the result of different states of the three-dimensional world, leaving the observer uncertain between these different states.

Another common source of uncertainty is randomness in the way an earlier state in a generative model generates a later state. One important type of such randomness pertains to the input the observer receives, such as the random pixel noise that corrupts the image of a grating in Fig. 1. An observer is left uncertain when I under-determines s because of noise.

Another important type of randomness is internal to the observer. The firing behavior of neurons is driven only partially by the signal they receive and partially by further random factors[45]. One and the same neural response r can be caused by different input states I, which in turn depend on different world states s. An actual observer does not infer s from I but from r; there is generally more uncertainty about s given r than given I owing to the additional uncertainty about I given r (Fig. 1c).

The extent to which randomness increases uncertainty depends on the amount of data available to the observer. For instance, when there is a fixed amount of pixel noise, the orientation task in Fig. 1 is easier for images with more pixels. Similarly, uncertainty about s decreases when evidence can be accumulated across multiple sensory inputs I, and it is large when only little evidence about s is observed, such as when s changes across time[11,46–49].

Studying representations of uncertainty with correlational and code-driven approaches

Empirical studies on the neural representation of uncertainty differ along various dimensions, such as the recording technique used, the species and the tasks. Here we propose a distinction that reflects a difference in approaches to uncertainty of distinct research communities, one rooted in cognitive neuroscience and the other in theoretical neuroscience. At the methodological level, this distinction is based on whether assumptions about neural coding of world states are used to study uncertainty.

The correlational approach to the representation of uncertainty does not make any (explicit) assumption about a neural code for representing s. Instead, researchers use a proxy for the brain's uncertainty about s that they derive from the input and behavior (denoted $u(I)$ and $u(b)$, respectively). This proxy then guides the search for parts of the brain whose activity r covaries with it. This search is loosely constrained by assumptions about the neural code of u. For instance, in a study that measures r with fMRI, a relation between u and r is typically tested for every single voxel in the brain.

The code-driven approach, by contrast, makes strong assumptions about the neural code of the representations of the latent world state s and the associated uncertainty u. On the basis of those assumptions, researchers can read out u from neural activity r. The model is tested by relating the uncertainty derived from the neural activity (denoted $u(r)$) with estimates of uncertainty derived from the sensory input ($u(I)$) or the behavioral response b.

Prototypical examples of the two approaches

We illustrate the correlational approach with an fMRI study from Vilares and colleagues[50]. In this study, participants were presented on each trial with an input I consisting of a dot-cloud sampled from a Gaussian distribution whose mean location was the latent state s that participants had to report. The experimenters used two dispersion levels for the dot-cloud (that is, variance of the Gaussian distribution), small and large, to manipulate the participant's uncertainty about s. The authors asked where the input-related uncertainty $u(I)$ (using the dispersion of the dot-cloud as a proxy) was represented in the brain, and stressed the absence of strong hypothesis: "it remains unknown whether uncertainty is represented along the sensorimotor pathway or within specialized brain areas outside this pathway." To detect and localize a representation of uncertainty, they regressed $u(I)$ against the fMRI signal in each voxel throughout the brain. They found that fMRI activity positively correlated with $u(I)$ in the early visual cortex.

As an illustration of the code-driven approach, consider a study from Geurts and colleagues[51], who presented participants with an input I consisting of an oriented grating (the orientation is the latent state s) with low contrast. Participants reported both the orientation of the grating and their uncertainty. The authors made very specific assumptions about the neural representation of s (the orientation) in the early visual cortex: "The model assumes that, across trials, voxel activity follows a multivariate normal distribution around the voxel's tuning curve for orientation." When fitted to the fMRI data, this model characterizes the probability of an activity pattern r given the orientation s. Using probabilistic inference, this model can be inverted to estimate the probability distribution of s given the observed r (Fig. 1), and the uncertainty about s inferred from r, $u(r)$, can be measured as the standard deviation of this distribution. To test whether it is indeed a neural representation of uncertainty about s, the authors regressed across trials $u(r)$ against $u(b)$, the uncertainty reported by participants, and they found a significant positive effect.

Estimates of uncertainty derived from the input and behavior

The above prototypical examples illustrate that the two approaches use some estimate $u(I)$ and $u(b)$ of uncertainty derived from the input or behavior and provide specific examples of such estimates; below we provide a broader range of examples. In general, the same estimates of $u(I)$ and $u(b)$ are available for either approach. What distinguishes the two approaches is the way those estimates are used: as a proxy for the brain's uncertainty to search for neural representations of this uncertainty in the correlational approach, and as a check of the neural readout of uncertainty in the code-driven approach.

Different aspects of behavior are used to derive estimates of uncertainty. In some studies, $u(b)$ is the uncertainty reported by human participants (for example, ratings or confidence judgments[46,47,51–60]). In other studies, researchers infer $u(b)$ by assuming that uncertainty regulates some specific aspects of participants' behavior[58,61], such as how fast to respond[62,63], how long to wait for a reward[30,64,65], whether to opt out of a bet[31,32,34,66–68], the variability of behavioral reports[57,69] or the relative weight between prior information and the current input I[50,70].

Different methods also exist to estimate uncertainty from the input. Some researchers use ideal observer models (see above), which are useful to quantify uncertainty across a wide variety of task structures, notably when the relation between u and I is complex (for example, in sequential learning)[7,46,47,71–73]. Other models used to estimate $u(I)$ go beyond the task-based generative model and incorporate assumptions about the decomposition of I into specific features by sensory systems[74–76]. Another method eschews the use of generative models of I by relying on simple qualitative relationships that exist between u and specific aspects of I (for example, pixel noise or contrast in the oriented grating example, or the fact that humans are more uncertain about oblique than cardinal orientations); these are used as crude estimates of uncertainty[30,31,69,77].

Assumptions about the neural code

The correlational approach seeks to relate the uncertainty proxy $u(I)$ or $u(b)$, to r (Fig. 2). In practice, researchers test for this relationship by different means, such as correlation, multiple linear regressions or multivariate pattern decoding (for example, with support-vector machines; Box 2). Each method implicitly makes assumptions about the neural code of u (for example, linearity in the case of correlation) but the choice of a method tends to be motivated more by convenience (the use of standard tools that capture simple statistical relationships between u and r) than strong hypotheses about the code. Depending on the method, the strength of this relationship is measured as a correlation coefficient, the significance of regression weights[78], the cross-validated decoding accuracy or the fraction of explained variance[79,80].

By contrast, the code-driven approach makes assumptions (in the form of a 'neural' generative model) about how r represents s. We present two families of such models[81]: those that posit a neural code for s in which r encodes a likelihood function $\mathcal{L}(s;r) = p(r|s)$ (broadly referred to as a probabilistic population code) and those that posit a neural code for s in which r corresponds to samples from a posterior distribution over s given the observed I, $p(s|I_{\text{observed}})$ (sampling-based code). Note that those two approaches are not necessarily contradictory[82]. Other models relevant for the study of uncertainty exist[3,83–87], but they have so far received less attention from experimenters.

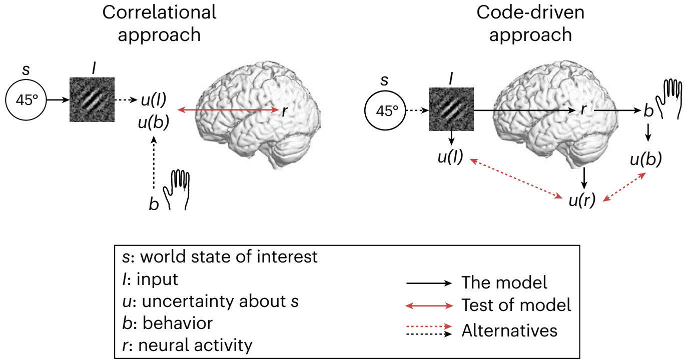

Fig. 2 | Comparison of the correlational and code-driven approaches. In both approaches, the participant is provided with an input *I* (here, an image of a grating) that is informative about a particular world state *s* (here, the orientation of the grating), the participant may provide a response (denoted by the hand) and researchers use some estimates of the brain's uncertainty *u* about *s*. Those estimates are derived from *I* itself (denoted *u*(*I*), based on the modeling of *p*(*I*|*s*) or a simple qualitative relationship between *u* and *I*) or *b* (denoted *u*(*b*), based on the participant' report of *u* or aspects of their choices that should depend on *u*). In the correlational approach, researchers test for a relationship (red arrow) between the proxy *u*(*I*) or *u*(*b*) and neural activity *r*, without making assumptions about the relationship between *s* (or *I*) and *r*. By contrast, the code-driven approach assumes a specific neural code for *s* (through *I*), denoted by the black arrow from *I* to *r*. On the basis of these assumptions, researchers can read out *u*(*r*), the uncertainty about *s* given the observed *r*; see black arrow from *r* to *u*(*r*). In practice, *u*(*r*) is obtained by inverting a neural likelihood function $\mathcal{L}(s;r_{\text{observed}}) = p(r_{\text{observed}}|s)$ in a probabilistic population code or by measuring the standard deviation of *r* in a sampling-based code. The validity of the code-based approach and the neural readout of uncertainty is evaluated by testing for some relationship between *u*(*r*) and either *u*(*b*) or *u*(*I*).

In probabilistic population codes, researchers formalize the uncertainty *u* conveyed by the neural activity observed on a particular trial, r_{observed}, as the posterior distribution $p(s|r_{\text{observed}})$. This posterior is derived using Bayes' rule from a neural likelihood function $\mathcal{L}(s;r_{\text{observed}})$ (and a prior over *s*, but this aspect is typically obviated by assuming non-informative priors). The construction of $\mathcal{L}(s;r)$ can be more or less data driven. The dominant approach in the literature is strongly theory driven[88,89]. An influential example in the sensory domain posits that neurons have a stereotyped mean response to the input (known as their tuning curve) and some variability corresponding to the exponential family of distributions (for example, Poisson distributions). Together with a few other assumptions, the log of $\mathcal{L}(s;r)$ becomes linear with respect to *r* and uncertainty about *s* is proportional to the average neural activity on a given trial[90–92]. By contrast, more data-driven approaches require fewer assumptions and estimate $\mathcal{L}(s;r)$ from the data itself. With the advent of large datasets and machine learning tools, even arbitrary shapes of $\mathcal{L}(s;r)$ can be estimated[35]. With a smaller amount of data, further constraints are needed about the shape of $\mathcal{L}(s;r)$, for example, assumptions about specific covariance matrices or noise distributions[51,57,69,70].

In sampling-based codes, the neural activity is assumed to represent *s* in terms of samples stochastically drawn from the posterior distribution $p(s|I_{\text{observed}})$[2,4,76,93]. Biologically plausible neural network models have been proposed for such a sampling process[94]. Under such a code, *u* is reflected in the spread of the distribution of *r* (for example, the standard deviation) across time or across neurons.

To summarize, studies within the correlational and code-driven approaches to studying uncertainty differ across many dimensions (the recording techniques; whether mathematical models, and *I* or *b*, are used to estimate *u*). The key difference is whether assumptions about the neural code of *s* are used to relate *r* to the uncertainty about *s*.

BOX 2

Relation to encoding and decoding approaches

Encoding and decoding are widely used notions in neuroscience[79,81,88,125–129]. The encoding approach models *r* as a function of some task-related quantity *x* (typically s or *I*, or *u*(*I*) or *u*(*b*) in the context of uncertainty); the decoding approach models *x* as a function of *r*. The relationship between encoding or decoding and correlational or code-driven approaches is multifaceted, notably because, in practice, different implementations of encoding and decoding exist in the domains of data analysis (machine learning) and theoretical neuroscience.

In theoretical neuroscience, encoding and decoding models are expressed in terms of conditional probabilities[88,129,130]. An encoding model corresponds to *p*(*r*|s) (and thus the neural likelihood function $\mathscr{L}$(s;*r*)), whereas a decoding model corresponds to *p*(s|*r*) (and thus potentially captures the brain's uncertainty about s). It is possible to obtain *p*(s|*r*) from *p*(*r*|s) together with a prior probability *p*(s) using Bayes' rule, which indicates that encoding and decoding models are related but not equivalent since a prior is also involved. The essence of probabilistic population codes is to model L(s;*r*) (encoding) and use it to obtain *p*(s|*r*) (decoding); thus, both encoding and decoding can be related to a code-driven approach to uncertainty.

By contrast, encoding and decoding in the data analysis domains[79,128] and machine learning applied to neuroscience[126,127,131] typically do not involve conditional probabilities between s and *r*. In this domain, an encoding model typically corresponds to first assuming a deterministic mapping (often highly nonlinear) from s (or *I*) to a list of latent features, and then testing for a relation to *r* by means of a (multiple) linear regression of the latent features onto *r* (linearizing encoding models[128]). Some studies that follow the correlational approach to uncertainty use this encoding method because they treat the uncertainty *u*(*I*) as a latent feature of the input *I*, and regress *u*(*I*) onto *r*. The same is true of sampling-based code studies except that they consider the variability of *r* rather than *r* itself. Regarding decoding models, they typically correspond to classifiers (for example, linear discriminant analysis or support-vector machine) that are trained to obtain s (or *I*) from *r*. This method is also used by some studies that follow the correlational approach, with the twist that the classifier is trained to obtain *u*(*I*) or *u*(*b*) instead of s itself.

Some applications of encoding or decoding are hybrid and use both a decomposition of s (or *I*) into some latent features that are regressed onto *r*, and conditional probabilities to model p(*r*|s) using the residuals of this regression. Some researchers in the code-driven approach used such a method to parametrize a probabilistic encoding model *p*(*r*|s) and then decode the uncertainty about s given *r*[51,57,69,70].

General criteria for the evaluation of claims about representations of uncertainty

We now turn to the evaluation of claims about the neural representation of uncertainty that can be found in either the correlational or code-driven approach. We propose to do so by applying criteria that are generally used in neuroscience to support claims about representations: sensitivity, specificity, invariance and functionality (Box 3).

BOX 3

General criteria for neural representations

Below we list criteria generally used in neuroscience to establish claims about representations (although they do not always appear under the labels we propose).

Sensitivity: r is sensitive to a feature x if changes in r are related to changes in x. For instance, a neuron is sensitive to the orientation of a bar if different activity patterns are recorded when different orientations are presented[132].

Specificity: r represents x specifically with respect to another feature y if changes in r are related to changes in x even when controlling for y. This criterion enables researchers to test that r is related to x indeed, and not spuriously so because of another feature y that is related to x (in that case, y is termed a confounding variable)[133]. For instance, uncertainty about orientation depends on the image contrast: a neural representation of the uncertainty about orientation therefore ought to be sensitive to contrast. However, to be a representation of uncertainty per se rather than contrast, r should reflect uncertainty even when the image contrast is kept fixed.

Invariance: The representation of x by r is invariant to y if changes in r are not related to changes in y when controlling for x. This criterion enables the researcher to test that r is related to x because r does not change when a feature y unrelated to x changes. For instance, the representation of orientation (our x) by V1 neurons (our r) is not invariant to position because different r are observed for a given orientation when changing position in the visual field[132]. By contrast, the representation of object identity in the inferotemporal cortex is invariant to the position and orientation of objects[116,134].

Functionality: r is functional as a representation of x if it causes a behavioral response b, for example, a report of the perceived value of x[135,136]; or, in the case of uncertainty, a decision that weighs sources of information by their respective uncertainty[50]. One can test for functionality with criteria analogous to the ones presented above; yet rather than testing for a dependence of r on x, these functionality analogs test for a dependence of b on r.

Claims about representations, ultimately, have to be claims about the causal structure of information processing in the brain. Nonetheless, we express the criteria in terms of information theoretic relationships between variables rather than in causal terms because researchers often use correlative (not causal) methods. Previous studies on uncertainty have used correlation[29,51,69], differences between conditions[31,32,50], linear regression[46,47,49,55] or decoding[55,103,107,137,138] to establish representations of uncertainty.

In practice, the criteria are evaluated in a graded manner such that they can be more or less satisfied (sensitivity might, for instance, be measured in terms of the strength of the correlation between r and x). Moreover, while linear relationships such as those illustrated in Fig. 3 are often simplest to understand and most common in experimental studies, many models, especially code-driven ones, posit nonlinear relationships.

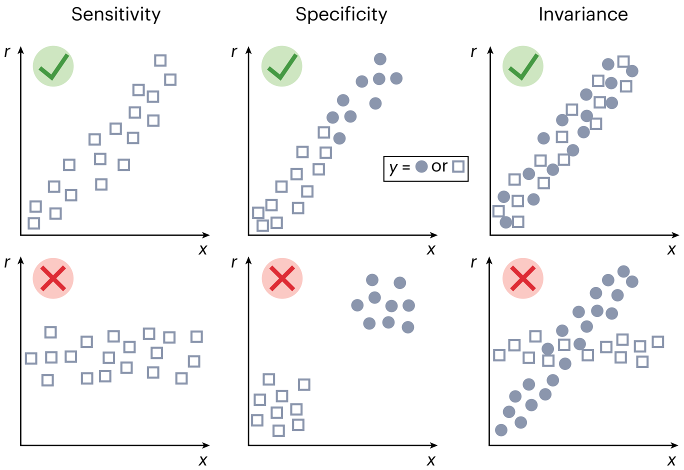

Fig. 3 | Empirical criteria for neural representation. We consider three criteria (sensitivity, specificity and invariance) that are used in neuroscience to test whether neural activity r represents a feature of interest x (like uncertainty, derived either from some input to the participant, $u(I)$, or from their behavior, $u(b)$), either by itself or in comparison to another feature y. The top row shows examples that pass a given criterion, and the bottom row shows examples that fail. In contrast to the correlational approach, the code-driven approach is interested in the uncertainty derived from r, $u(r)$, rather than r itself; in that case, $u(r)$ replaces r in the above graphs. Also note that the other feature y is categorical here, but it could also be continuous.

How do the two approaches compare in terms of testing general criteria?

Sensitivity. Testing for sensitivity is the starting point of the correlational approach because it aims to identify whether some r in the brain is sensitive to a proxy $u(I)$ for uncertainty (for $u(b)$, see 'Functionality'). The higher the sensitivity, such as the strength of correlation or decoding accuracy, the more plausible it is that a given r is a representation of u. The use of better proxies for uncertainty also makes tests of sensitivity more convincing. Some studies that follow the correlational approach address only sensitivity, especially when they are among the first of their kind or when uncertainty is not central in the study[48,49,95].

The code-driven approach tests whether a neural readout of uncertainty $u(r)$ is sensitive to $u(I)$. In this case, the vertical axis in Fig. 3 is $u(r)$, not r as in the correlational approach, although in both cases $x = u(I)$. To illustrate, $u(r)$ was shown to be sensitive to aspects of I that impact uncertainty, such as the image contrast[35,76,94], whether orientation is closer to cardinal axes[69], or the presence of a higher-level context[96]. Although either a probabilistic population code or a sampling-based code was used to derive $u(r)$ in those example studies, a prominent difference is that the sensitivity test is more central in studies that use a sampling-based code. In the case of the probabilistic population code, sensitivity sometimes appears as a side point[35] or is even not reported[70].

Specificity. In both approaches, researchers correlate (or use more elaborate analyses) $u(I)$ or $u(b)$ to r or $u(r)$. Testing for correlation is vulnerable to the problem of confounding variables: r may not represent u but the aspect of I or b from which $u(I)$ or $u(b)$ has been derived, such as contrast in the orientation task. It is still possible to test for specificity if several features of I or b are related to u. In that case, the specificity of r to u with respect to each feature in isolation can be tested by holding each feature fixed and testing for the dependence of r on other features of I or b[27,74,77]. For example, Dekleva and colleagues[27] manipulated uncertainty about the direction of reaching in a motor task through the current trial's cue and the cue history, and Bang and colleagues[77] manipulated uncertainty about the direction of motion by changing the strength of motion evidence and the distance to the category boundary. In both studies, r continued to track $u(I)$ when either feature was kept fixed. Some aspects of I can be artifactually correlated with u. For instance, uncertainty about local features in an image is expected to decrease when they are embedded in a higher-level structure; one can test for this effect while controlling for the spectral content of the image, which is often confounded with the presence of high-level structure[96]. Some previous studies include tests for confounding variables such as reaction times[97], attention[97], exploration[7,98] and task difficulty[30,31,64,66,77].

When $u(I)$ is not derived from a simple feature of the input but from a more complex model, such as an ideal observer model, several confounding variables might still undermine the specificity of r to u. For instance, in the context of sequential learning, u is often negatively correlated with recently surprising outcomes[46,47,99]. Confounding variables of uncertainty about the present world state also include constructs presented in other studies such as the likelihood of a change point[11], expected uncertainty[100], total uncertainty[9,71], outcome uncertainty[46,47,101] and expected reward[102–105].

Invariance. Some researchers following the correlational approach have tested for invariance. For instance, Michael and colleagues[106] used a categorization task and inputs with two features: shape and color. The relevant feature used for the categorization task changed across trials and a common neural representation r of the categorization uncertainty was found in both conditions. In the shape condition, r tracked the uncertainty related to shape, not color (and vice versa in the color condition), demonstrating that r coded for uncertainty beyond these low-level features. Using a similar logic, Lebreton and colleagues[56] found a general neural representation of the uncertainty associated with estimating the value of paintings, objects and prospects. Other researchers have tested invariance with respect to sensory modality[47,64,107].

Invariance is rarely tested in the code-driven approach. Orban et al. found that $u(r)$ (in this case, the variability of the membrane potential) was sensitive to some $u(I)$ (the contrast of a grating) and tested for invariance with respect to the orientation of the grating. They reported 'mild modulations' by orientation when $u(I)$ was kept fixed[76].

Functionality. The correlational approach can use $u(b)$ instead of $u(I)$ as just another proxy for u. In that case, functional sensitivity (tested as a relation between r or $u(r)$ and $u(b)$) is not fundamentally different from the sensitivity test presented above (based on $u(I)$). However, $u(I)$ and $u(b)$ are unlikely to be equivalent and it is unclear whether one is a better proxy for the brain's u than the other, because the ability of participants to introspect u may be limited[108,109] and additional processes may intervene between u and choices or reports based on u[60,110]. Instead of using $u(b)$ in place of $u(I)$ as just another proxy for u, more compelling evidence of functionality comes from correlational studies that combine $u(I)$ and $u(b)$. For instance, a neural representation r of u based on $u(I)$ is identified first and then some relation between this r and b is sought. Some studies reported correlations between neural representations of $u(I)$ and the reported uncertainty[46], trial-to-trial variability during reaching[27] and learning[11]. Other studies reported correlations across participants between neural representations of $u(I)$ and aspects of behavior that should in principle depend on uncertainty, such as risk attitude[111], exploration[9,71] and prior-likelihood combination[50,112].

In the context of the code-driven approach with a probabilistic population code, the functionality criterion often plays a key role. For instance, van Bergen and colleagues[69] showed that $u(r)$, the uncertainty about the orientation of a grating inferred from V1 fMRI activity, correlated with the variability of orientation reports. The same group also found that this $u(r)$ correlated across trials with both the uncertainty reported by participants[51] and the strength of sequential effects in their perceptual decisions[70]. Walker and colleagues[35] found that the uncertainty read out from V1 in monkeys accounted for their decisions in an uncertainty-based categorization task.

By contrast, functionality is less often tested in studies that assume a sampling-based code. An exception is a 2016 study by Haefner and colleagues[113], which reported that the structure of covariance among artificial neurons in a network reflected the uncertainty about the task-relevant orientation during visual categorization in a way that correlated with performance in the task.

Functionality can be coupled with specificity. Such a test assesses whether r or $u(r)$ still correlates with $u(b)$ when all features of the input I are held constant. Uncovering such an effect suggests that it is indeed uncertainty, and not any confounding feature of the input, that is represented and used for behavior. The code-driven approach lends itself to such a test because researchers can read out u from r, making it possible to find correlation between $u(r)$ and $u(b)$ when I is fixed[35,69]. The correlational approach does not seem suited for this test when it uses $u(I)$. But, some analyses inspired by this test are still informative; for instance, McGuire and colleagues used $u(I)$ from an ideal observer to identify a neural representation r of $u(I)$; they then regressed $u(I)$ out of r (which is analogous to keeping the effect of I fixed) and showed that r still correlated with some aspect of behavior[11]. Showing that r or $u(r)$ is sensitive to $u(I)$, and to $u(b)$ on top of the effect of $u(I)$, indicates that the way the brain computes uncertainty based on the input deviates from the model assumed to derive $u(I)$, or that other processes (for example, internal noise, attention[114] or biases) intervene in the computation of uncertainty or its use in behavior.

How do the two approaches compare in terms of satisfying general criteria?

Supplementary Table 1 reports examples of previous studies on the neural representation of uncertainty. The uncertainty that characterizes representations of the world by an observer is often distinguished from the uncertainty about the outcome of a process[95,101] and from decision confidence[60]. Yet, the corresponding studies often test the same criteria and face similar methodological problems; therefore, we include all of them in this table.

Supplementary Table 1 shows how the different approaches (correlational, code driven) and the different types of uncertainty compare in terms of satisfying the criteria. The criteria apply to both approaches, but with some differences: functionality is currently a strong point of the code-driven approach (when it relies on probabilistic population codes, not on sampling-based codes) in comparison to the correlational approach, whereas invariance, and to a lesser extent specificity, are more often tested in the correlational approach than the code-driven approach.

Caveats of current approaches and future directions

We now summarize the potential and limitations of the two approaches.

Comparing correlational and code-driven approaches

Assumptions about neural codes. We based our distinction between code-driven and correlational approaches on whether assumptions are made about the neural coding of the world state s and the accompanying uncertainty. This methodological difference has conceptual implications. The code-driven approach studies neural representations in which a neural population r jointly represents both a world state s and uncertainty u about s. The correlational approach does not require such joint representations and, therefore, it can identify representations of u that are not colocalized with the representation of s[50]. It has been proposed that some brain regions could be specialized in the representation and processing of uncertainty[58,101,108,115]; such brain regions can be identified with the correlational approach but not with the code-driven approach in its current form.

This restriction to joint representations is likely to explain different findings between the two approaches. The code-driven approach more often identifies representations of u in sensory regions such as the early visual cortex, which are well known for representing visual features[35,51,69,70,75,76], whereas the correlational approach often identifies representations of u in regions that are further from sensory input and its representation, closer to the decision or reporting mechanisms, in subcortical structures[49,50,97,112], prefrontal cortex[11,46,49,50,56,71,98,106], parietal cortex[11,47,49,71,98,106] or temporal cortex[71,111].

The two approaches also differ in assumptions about the complexity of the neural code of uncertainty. In the correlational approach, codes are usually assumed to be linear (monotonic changes in average activity as a function of uncertainty); representations with such linear codes have been termed explicit representations[116,117]. Studies following the code-driven approach are open to nonlinear computations, which are often used to derive u from r, for example, when reading out the standard deviation of the decoded distribution[51,69,70], or the standard deviation of neural activity[76] or when using artificial neural networks[35].

Specificity. Both approaches use external estimates of uncertainty $u(I)$ and $u(b)$ and thus are susceptible to confounding factors. The code-driven approach has an elegant method to demonstrate functional specificity with respect to the features of the input I, namely by testing whether $u(r)$ makes a difference to the behavioral response b, even while controlling for I[35]. Yet, concerns about specificity remain even in this case because representations of uncertainty may still be confounded by behavioral features or processes internal to the brain, such as attention.

Functionality. Behavioral data are used with substantial heterogeneity in both approaches. In the code-driven approach in particular, it is striking that $u(b)$ is much more prominent than $u(I)$ in studies using probabilistic population codes, and that the converse is true in those using sampling-based codes. Interestingly, tests based on $u(b)$ have fewer degrees of freedom and thus seem more stringent in the probabilistic population code-driven approach than in many studies following the correlational approach. To illustrate, the test is passed in the code-driven approach only if $u(r)$ correlates with $u(b)$, the participant's report of uncertainty[51], whereas it is passed in the correlational approach if r correlates with $u(b)$ in at least one of the many brain regions under investigation.

Origins of uncertainty. Externally generated uncertainty $u(I)$ derives from ambiguity or noise in the generation of the sensory input I and can be estimated with an ideal observer model. Internally generated uncertainty depends on neural noise or limitations and errors in information processing; it can only be estimated from behavioral responses $u(b)$ or neural activity $u(r)$; $u(b)$ and $u(r)$ also track external sources of uncertainty. Studies following the correlational approach that focus on $u(I)$ are restricted to externally generated uncertainty. However, by including $u(b)$ in the analysis, correlational studies can also account for internally generated uncertainty. Because they read the uncertainty $u(r)$ directly from the brain state r, code-driven models are especially well suited to studying internally generated uncertainty. Interestingly, studies using sampling-based codes currently assume that those internal sources of uncertainty (noise) are negligible and that $u(r)$ correspond to the uncertainty optimally computed from I[76]; by contrast, studies using probabilistic population codes stress the importance of internal sources of uncertainty[92,118].

Setting goals for future research

Given that the code-driven and correlational approaches have different limitations and advantages, they could be used in synergy. One possibility is to leverage our knowledge of early sensory cortices to have a neural readout of uncertainty $u(r)$ about s in a perceptual task using the code-driven approach, and then use $u(r)$ as input to the correlational approach to unravel other parts of the brain that could represent this uncertainty without requiring that they represent s itself. Such combined analysis would reconcile the fact that the representation of uncertainty can be colocalized with the representation of the feature s that it characterizes while also being detached from it by downstream computation. Some studies have already started to reduce the gap between code-driven and correlational approaches. The study by Geurts et al.[51] that we used as a prototypical example of the code-driven approach also used the correlational approach and found fMRI correlates of $u(r)$ in the prefrontal cortex.

Understanding how the brain extracts and uses uncertainty can also be achieved by further investigation of the functional aspect of representations. If uncertainty is used only in a given context (for example, uncertainty about color, not shape, is relevant for color-based categorization[106]) or for different goals (for example, to guide the decision to wager[31] or to update prior estimates[47]), then some aspects of its representation are expected to change. Manipulating the task relevance of uncertainty is thus a promising avenue to explore the function of the representation of uncertainty. In particular, it would be useful to distinguish representations of uncertainty that are automatic and occur independently of task demands from those that are task dependent[119].

We have stressed that uncertainty can be about different things (for example, orientation of a grating[35], color[106], the next outcome[103,120–122] and probability of an event[47]) and have multiple origins (for example, prior knowledge and current input). Whether representations of uncertainty are invariant to the origin of this uncertainty, and invariant to what uncertainty is about, remains a largely open question. A related methodological concern, in particular for the code-driven approach, is that r may actually not represent the world state s of interest to the researcher but some other feature z; substantial difference between the uncertainty about s and z given I will undermine the code-driven approach. For instance, r in V1 may represent not orientation but instead the intensity of a specific set of image elements present in I[82].

As the field matures, a switch from single-model testing to the comparison of different models (for example, generative models of the observer and the brain used to infer $u(I)$ and $u(r)$; linear versus nonlinear neural codes for u in the correlational approach; code-driven approaches that disentangle the representations of s and u) would be valuable to narrow down the neural codes of uncertainty. Because sampling-based codes and probabilistic population codes focus on encoding and decoding, respectively, they could also be combined to model processes from input to behavior.

Manipulating prior expectations could help to tackle the pervasive issue of specificity: posterior uncertainty depends on both the current input and the prior, but most studies focus on the former. Manipulating priors enables researchers to partly de-correlate posterior uncertainty from the current input. Some previous studies manipulated priors[50,112] but with the aim of comparing the encoding of the prior and current likelihood. Beyond the methodological interest regarding specificity, systematic manipulation of priors (as in previous behavioral studies[123]) would also be useful to study at which stage prior and current uncertainties are combined in the brain when processing the current input, and to compare empirically probabilistic population codes and sampling-based codes.

In conclusion, we propose that current studies on the neural representation of uncertainty can be distinguished as code-driven versus correlational approaches on the basis of whether they rely on assumptions about the neural code of some world state and the accompanying uncertainty. This distinction results in the identification of potentially different types of representation of uncertainty that may be colocalized with, or separated from, the representation of the corresponding world state. Empirical conclusions from both approaches can be assessed with the same set of general criteria, but there is currently an emphasis on different criteria across studies. Because the two approaches differ in the assumptions they require and the types of finding they uncover, there is great potential for them to be used synergistically.

References

1. Ballard, D. H. *Brain Computation as Hierarchical Abstraction* (MIT Press, 2015).

2. Hoyer, P. O. & Hyvärinen, A. Interpreting neural response variability as Monte Carlo sampling of the posterior. In *Advances in Neural Information Processing Systems* 293–300 (2002).
An influential article that proposed that neural activity could be explained with a sampling-based code.
3. Knill, D. C. & Pouget, A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* **27**, 712–719 (2004).
4. Lee, T. S. & Mumford, D. Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **20**, 1434–1448 (2003).
5. Ma, W. J. & Jazayeri, M. Neural coding of uncertainty and probability. *Annu. Rev. Neurosci.* **37**, 205–220 (2014).
6. Bach, D. R. & Dolan, R. J. Knowing how much you don't know: a neural organization of uncertainty estimates. *Nat. Rev. Neurosci.* **13**, 572–586 (2012).
7. Muller, T. H., Mars, R. B., Behrens, T. E. & O'Reilly, J. X. Control of entropy in neural models of environmental state. *eLife* **8**, e39404 (2019).
8. Rushworth, M. F. S. & Behrens, T. E. J. Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat. Neurosci.* **11**, 389–397 (2008).
9. Tomov, M. S., Truong, V. Q., Hundia, R. A. & Gershman, S. J. Dissociable neural correlates of uncertainty underlie different exploration strategies. *Nat. Commun.* **11**, 2371 (2020).
10. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).
11. McGuire, J. T., Nassar, M. R., Gold, J. I. & Kable, J. W. Functionally dissociable influences on learning rate in a dynamic environment. *Neuron* **84**, 870–881 (2014).
12. Meyniel, F., Schlunegger, D. & Dehaene, S. The sense of confidence during probabilistic learning: a normative account. *PLoS Comput Biol.* **11**, e1004305 (2015).
13. O'Reilly, J. X. Making predictions in a changing world—inference, uncertainty, and learning. *Front. Neurosci.* **7**, 105 (2013).
14. Kersten, D., Mamassian, P. & Yuille, A. Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304 (2004).
15. Qamar, A. T. et al. Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proc. Natl Acad. Sci. USA* **110**, 20332–20337 (2013).
16. Zhou, Y., Acerbi, L. & Ma, W. J. The role of sensory uncertainty in simple contour integration. *PLoS Comput. Biol.* **16**, e1006308 (2020).
17. Alais, D. & Burr, D. The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* **14**, 257–262 (2004).
18. Deroy, O., Spence, C. & Noppeney, U. Metacognition in multisensory perception. *Trends Cogn. Sci.* **20**, 736–747 (2016).
19. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
20. Trommershäuser, J., Kording, K. & Landy, M. S. *Sensory Cue Integration* (Oxford Univ. Press, 2011).
21. Todorov, E. Optimality principles in sensorimotor control. *Nat. Neurosci.* **7**, 907–915 (2004).
22. Trommershäuser, J., Maloney, L. T. & Landy, M. S. Decision making, movement planning and statistical decision theory. *Trends Cogn. Sci.* **12**, 291–297 (2008).
23. Flavell, J. H. & Wellman, H. M. in *Perspectives on the Development of Memory and Cognition* (eds. Kail, R. V. Jr & Hagen, J. W.) 3–33 (L. Erlbaum, 1977).
24. Koriat, A., Sheffer, L. & Ma'ayan, H. Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *J. Exp. Psychol.* **131**, 147–162 (2002).
25. Rademaker, R. L., Tredway, C. H. & Tong, F. Introspective judgments predict the precision and likelihood of successful maintenance of visual working memory. *J. Vis.* **12**, 21 (2012).
26. Yoo, A. H., Acerbi, L. & Ma, W. J. Uncertainty is maintained and used in working memory. *J. Vis.* **21**, 13 (2021).
27. Dekleva, B. M., Ramkumar, P., Wanda, P. A., Kording, K. P. & Miller, L. E. Uncertainty leads to persistent effects on reach representations in dorsal premotor cortex. *eLife* **5**, e14316 (2016).
28. Devkar, D., Wright, A. A. & Ma, W. J. Monkeys and humans take local uncertainty into account when localizing a change. *J. Vis.* **17**, 4 (2017).
29. Fiorillo, C. D. Discrete coding of reward probability and uncertainty by dopamine neurons. *Science* **299**, 1898–1902 (2003).
30. Kepecs, A., Uchida, N., Zariwala, H. A. & Mainen, Z. F. Neural correlates, computation and behavioural impact of decision confidence. *Nature* **455**, 227–231 (2008).
31. Kiani, R. & Shadlen, M. N. Representation of confidence associated with a decision by neurons in the parietal cortex. *Science* **324**, 759–764 (2009).
32. Komura, Y., Nikkuni, A., Hirashima, N., Uetake, T. & Miyamoto, A. Responses of pulvinar neurons reflect a subject's confidence in visual categorization. *Nat. Neurosci.* **16**, 749–755 (2013).
33. Lak, A. et al. Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* **84**, 190–201 (2014).
34. Odegaard, B. et al. Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proc. Natl Acad. Sci. USA* **115**, E1588–E1597 (2018).
35. 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 (2020).
Example of the code-driven approach that uses a probabilistic population code estimated in a data-driven manner by means of an artificial neural network. The uncertainty derived from multiunit recordings accounts for the monkey choices.
36. Helmholtz, H. *Handbuch der Physiologischen Optik* (Leopold Voss, 1867).
37. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
38. Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. & Pouget, A. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron* **74**, 30–39 (2012).
39. Rahnev, D. & Denison, R. N. Suboptimality in perceptual decision making. *Behav. Brain Sci.* **41**, e223 (2018).
40. Iglesias, S. et al. Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron* **80**, 519–530 (2013).
41. Mathys, C. D. et al. Uncertainty in perception and the hierarchical gaussian filter. *Front. Hum. Neurosci.* **8**, 825 (2014).
42. Norton, E. H., Acerbi, L., Ma, W. J. & Landy, M. S. Human online adaptation to changes in prior probability. *PLOS Comput. Biol.* **15**, e1006681 (2019).
43. Barthelmé, S. & Mamassian, P. Evaluation of objective uncertainty in the visual system. *PLoS Comput. Biol.* **5**, e1000504 (2009).
44. Necker, L. A. Observations on some remarkable optical phænomena seen in Switzerland; and on an optical phænomenon which occurs on viewing a figure of a crystal or geometrical solid. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1**, 329–337 (1832).
45. Faisal, A. A., Selen, L. P. J. & Wolpert, D. M. Noise in the nervous system. *Nat. Rev. Neurosci.* **9**, 292–303 (2008).
46. Meyniel, F. Brain dynamics for confidence-weighted learning. *PLOS Comput. Biol.* **16**, e1007935 (2020).
47. Meyniel, F. & Dehaene, S. Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1615773114 (2017).

Example of a correlational approach that uses an ideal observer model of the input to derive uncertainty about a probability. The study reports fMRI correlates of this uncertainty distinct from correlates of confounding factors like unpredictability and surprise.

48. O'Reilly, J. X., Jbabdi, S., Rushworth, M. F. S. & Behrens, T. E. J. Brain systems for probabilistic and dynamic prediction: computational specificity and integration. *PLoS Biol.* **11**, e1001662 (2013).
49. Payzan-LeNestour, E., Dunne, S., Bossaerts, P. & O'Doherty, J. P. The neural representation of unexpected uncertainty during value-based decision making. *Neuron* **79**, 191–201 (2013).
50. Vilares, I., Howard, J. D., Fernandes, H. L., Gottfried, J. A. & Kording, K. P. Differential representations of prior and likelihood uncertainty in the human brain. *Curr. Biol.* **22**, 1641–1648 (2012).
    Example of correlational approach that used specific features of the input (scatter) as a proxy for uncertainty (about the location of a cloud of dots). The fMRI correlates of this uncertainty are distinct from prior uncertainty.
51. Geurts, L. S., Cooke, J. R. H., van Bergen, R. S. & Jehee, J. F. M. Subjective confidence reflects representation of Bayesian probability in cortex. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-021-01247-w (2022).
    Example of a code-driven approach that uses a probabilistic population code estimated in a data-driven manner by means of a generalized linear model. The uncertainty derived from fMRI activity correlates with subjective reports of uncertainty.
52. Adler, W. T. & Ma, W. J. Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLOS Comput. Biol.* **14**, e1006572 (2018).
53. De Martino, B., Fleming, S. M., Garrett, N. & Dolan, R. J. Confidence in value-based choice. *Nat. Neurosci.* **16**, 105–110 (2013).
54. Guggenmos, M., Wilbertz, G., Hebart, M. N. & Sterzer, P. Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *eLife* **5**, e13388 (2016).
55. Hebart, M. N., Schriever, Y., Donner, T. H. & Haynes, J.-D. The relationship between perceptual decision variables and confidence in the human brain. *Cereb. Cortex* https://doi.org/10.1093/cercor/bhu181 (2014).
56. Lebreton, M., Abitbol, R., Daunizeau, J. & Pessiglione, M. Automatic integration of confidence in the brain valuation signal. *Nat. Neurosci.* **18**, 1159–1167 (2015).
57. Li, H.-H., Sprague, T. C., Yoo, A. H., Ma, W. J. & Curtis, C. E. Joint representation of working memory and uncertainty in human cortex. *Neuron* **109**, 3699–3712 (2021).
58. Meyniel, F., Sigman, M. & Mainen, Z. F. Confidence as bayesian probability: from neural origins to behavior. *Neuron* **88**, 78–92 (2015).
59. Peirce, C. S. & Jastrow, J. On small differences in sensation. *Mem. Natl Acad. Sci.* **3**, 75–83 (1884).
60. Pouget, A., Drugowitsch, J. & Kepecs, A. Confidence and certainty: distinct probabilistic quantities for different goals. *Nat. Neurosci.* **19**, 366–374 (2016).
61. Kepecs, A. & Mainen, Z. F. A computational framework for the study of confidence in humans and animals. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 1322–1337 (2012).
62. Tzagarakis, C., Ince, N. F., Leuthold, A. C. & Pellizzer, G. Beta-band activity during motor planning reflects response uncertainty. *J. Neurosci.* **30**, 11270–11277 (2010).
63. Zylberberg, A., Fetsch, C. R. & Shadlen, M. N. The influence of evidence volatility on choice, reaction time and confidence in a perceptual decision. *eLife* **5**, e17688 (2016).
64. Masset, P., Ott, T., Lak, A., Hirokawa, J. & Kepecs, A. Behavior- and modality-general representation of confidence in orbitofrontal cortex. *Cell* **182**, 112–126 (2020).
    Studies decision confidence in rats using waiting times as a proxy for uncertainty and identifies a neural representation of decision confidence in the orbitofrontal cortex that passes the tests of sensitivity, specificity (with respect to the features of the input), invariance (to the sensory modality) and functionality (correlation with learning).
65. Schmack, K., Bosc, M., Ott, T., Sturgill, J. F. & Kepecs, A. Striatal dopamine mediates hallucination-like perception in mice. *Science* **372**, eabf4740 (2021).
66. Gherman, S. & Philiastides, M. G. Neural representations of confidence emerge from the process of decision formation during perceptual choices. *NeuroImage* **106**, 134–143 (2015).
67. Hampton, R. R. Rhesus monkeys know when they remember. *Proc. Natl Acad. Sci. USA* **98**, 5359–5362 (2001).
68. Middlebrooks, P. G. & Sommer, M. A. Neuronal correlates of metacognition in primate frontal cortex. *Neuron* **75**, 517–530 (2012).
69. van Bergen, R. S., Ma, W. J., Pratte, M. S. & Jehee, J. F. M. Sensory uncertainty decoded from visual cortex predicts behavior. *Nat. Neurosci.* **18**, 1728–1730 (2015).
70. van Bergen, R. S. & Jehee, J. F. M. Probabilistic representation in human visual cortex reflects uncertainty in serial decisions. *J. Neurosci.* **39**, 8164–8176 (2019).
71. Badre, D., Doll, B. B., Long, N. M. & Frank, M. J. Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron* **73**, 595–607 (2012).
    Example of a correlational approach that uses an ideal observer model of the learning process to infer uncertainty in a task. Findings show evidence of a functional role for uncertainty (here, in terms of exploration).
72. Stern, E. R., Gonzalez, R., Welsh, R. C. & Taylor, S. F. Updating beliefs for a decision: neural correlates of uncertainty and underconfidence. *J. Neurosci.* **30**, 8032–8041 (2010).
73. Sedley, W. et al. Neural signatures of perceptual inference. *eLife* **5**, e11476 (2016).
74. 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).
75. Hénaff, O. J., Boundy-Singer, Z. M., Meding, K., Ziemba, C. M. & Goris, R. L. T. Representation of visual uncertainty through neural gain variability. *Nat. Commun.* **11**, 2513 (2020).
76. Orbán, G., Berkes, P., Fiser, J. & Lengyel, M. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron* **92**, 530–543 (2016).
    Example of a code-driven approach that uses a sampling-based code and finds that neural variability (in spiking activity and membrane potential) changes along features of visual input related to uncertainty (for example, it quenches at the stimulus onset, decreases with contrast and aperture).
77. Bang, D. & Fleming, S. M. Distinct encoding of decision confidence in human medial prefrontal cortex. *Proc. Natl Acad. Sci. USA* **115**, 6082–6087 (2018).
78. Friston, K., Ashburner, J., Kiebel, S., Nichols, T. & Penny, W. *Statistical Parametric Mapping: the Analysis of Functional Brain Images* (Academic, 2007).
79. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410 (2011).
80. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of best practices for evidence for prediction: a review. *JAMA Psychiatry* **77**, 534–540 (2020).
81. Lange, R. D., Shivkumar, S., Chattoraj, A. & Haefner, R. M. Bayesian encoding and decoding as distinct perspectives on neural coding. Preprint at *bioRxiv* https://doi.org/10.1101/2020.10.14.339770 (2021).

82. Shivkumar, S., Lange, R., Chattoraj, A. & Haefner, R. A probabilistic population code based on neural samples. In *Advances in Neural Information Processing Systems* (eds. S. Bengio et al.) 31, 1–10 (MIT Press, 2018).
83. Barlow, H. B. Pattern recognition and the responses of sensory neurons. *Ann. N. Y. Acad. Sci.* **156**, 872–881 (1969).
84. Deneve, S. Bayesian spiking neurons I: inference. *Neural Comput.* **20**, 91–117 (2008).
85. Jazayeri, M. & Movshon, J. A. Optimal representation of sensory information by neural populations. *Nat. Neurosci.* **9**, 690–696 (2006).
86. Sahani, M. & Dayan, P. Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. *Neural Comput.* **15**, 2255–2279 (2003).
87. Sohn, H. & Narain, D. Neural implementations of Bayesian inference. *Curr. Opin. Neurobiol.* **70**, 121–129 (2021).
88. Dayan, P. & Abbott, L. F. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (MIT Press, 2005).
89. Pouget, A., Dayan, P. & Zemel, R. S. Inference and computation with population codes. *Annu. Rev. Neurosci.* **26**, 381–410 (2003).
90. Deneve, S., Latham, P. E. & Pouget, A. Reading population codes: a neural implementation of ideal observers. *Nat. Neurosci.* **2**, 740–745 (1999).
91. 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 (2012).
92. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432–1438 (2006).
Introduced the concept of probabilistic population code as the idea that the representation of probability distribution over a latent world state by a population of neurons, conferred by an internal model of neural variability, allows certain Bayesian computations to be implemented by simple neural operations.
93. 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).
94. 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).
Shows that an artificial neural network can be trained to emit spikes that correspond to samples from a posterior distribution of some feature of the input. Although not trained to do so, the artificial network shows dynamics similar to those of actual neural networks.
95. Bach, D. R., Hulme, O., Penny, W. D. & Dolan, R. J. The known unknowns: neural representation of second-order uncertainty, and ambiguity. *J. Neurosci.* **31**, 4811–4820 (2011).
96. Bányai, M. et al. Stimulus complexity shapes response correlations in primary visual cortex. *Proc. Natl Acad. Sci. USA* **116**, 2723–2732 (2019).
Example of a code-driven approach that uses a sampling-based code and shows that the covariance of neural activity in a population of neurons can be explained by hierarchical inference with a prominent impact of the image's higher-level features even in regions tuned to local features, such as the primary visual cortex.
97. Grinband, J., Hirsch, J. & Ferrera, V. P. A neural representation of categorization uncertainty in the human brain. *Neuron* **49**, 757–763 (2006).
98. Trudel, N. et al. Polarity of uncertainty representation during exploration and exploitation in ventromedial prefrontal cortex. *Nat. Hum. Behav.* **5**, 83–98 (2021).
99. Strange, B. A., Duggins, A., Penny, W., Dolan, R. J. & Friston, K. J. Information theory, novelty and hippocampal responses: unpredicted or unpredictable? *Neural Netw.* **18**, 225–230 (2005).
100. Tan, H., Wade, C. & Brown, P. Post-movement beta activity in sensorimotor cortex indexes confidence in the estimations from internal models. *J. Neurosci.* **36**, 1516–1528 (2016).
101. Hsu, M., Bhatt, M., Adolphs, R., Tranel, D. & Camerer, C. F. Neural systems responding to degrees of uncertainty in human decision-making. *Science* **310**, 1680–1683 (2005).
Presented a distinction between uncertainty about a latent feature and uncertainty about an outcome (referred to as ambiguity and risk, respectively, in behavioral economics), whose fMRI correlates are anatomically segregated in the human brain.
102. Monosov, I. E., Leopold, D. A. & Hikosaka, O. Neurons in the primate medial basal forebrain signal combined information about reward uncertainty, value, and punishment anticipation. *J. Neurosci.* **35**, 7443–7459 (2015).
103. Monosov, I. E. & Hikosaka, O. Selective and graded coding of reward uncertainty by neurons in the primate anterodorsal septal region. *Nat. Neurosci.* **16**, 756–762 (2013).
104. Preuschoff, K., Bossaerts, P. & Quartz, S. R. Neural differentiation of expected reward and risk in human subcortical structures. *Neuron* **51**, 381–390 (2006).
105. So, N. & Stuphorn, V. Supplementary eye field encodes confidence in decisions under risk. *Cereb. Cortex* **26**, 764–782 (2016).
106. Michael, E., de Gardelle, V., Nevado-Holgado, A. & Summerfield, C. Unreliable evidence: 2 sources of uncertainty during perceptual choice. *Cereb. Cortex* **25**, 937–947 (2015).
Example of a correlational approach that uses a categorization task based on either shape or color from trial to trial and identifies representations of uncertainty about the decision that are invariant to the perceptual feature (shape or color) on which a decision is based.
107. Nastase, S. A., Davis, B. & Hasson, U. Cross-modal and non-monotonic representations of statistical regularity are encoded in local neural response patterns. *NeuroImage* **173**, 509–517 (2018).
108. Fleming, S. M. & Daw, N. D. Self-evaluation of decision-making: a general Bayesian framework for metacognitive computation. *Psychol. Rev.* **124**, 91–114 (2017).
109. Zylberberg, A., Roelfsema, P. R. & Sigman, M. Variance misperception explains illusions of confidence in simple perceptual decisions. *Conscious. Cognition* **27**, 246–253 (2014).
110. Fleming, S. M. & Dolan, R. J. Effects of loss aversion on post-decision wagering: implications for measures of awareness. *Conscious. Cognition* **19**, 352–363 (2010).
111. Blankenstein, N. E., Peper, J. S., Crone, E. A. & van Duijvenvoorde, A. C. K. Neural mechanisms underlying risk and ambiguity attitudes. *J. Cogn. Neurosci.* **29**, 1845–1859 (2017).
112. Ting, C. -C., Yu, C. -C., Maloney, L. T. & Wu, S. -W. Neural mechanisms for integrating prior knowledge and likelihood in value-based probabilistic inference. *J. Neurosci.* **35**, 1792–1805 (2015).
113. Haefner, R. M., Berkes, P. & Fiser, J. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron* **90**, 649–660 (2016).
114. Rahnev, D. et al. Attention induces conservative subjective biases in visual perception. *Nat. Neurosci.* **14**, 1513–1515 (2011).
115. Schultz, W. et al. Explicit neural signals reflecting reward uncertainty. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **363**, 3801–3811 (2008).
116. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).

117. Kriegeskorte, N. & Diedrichsen, J. Peeling the onion of brain representations. *Annu. Rev. Neurosci.* **42**, 407–432 (2019).
118. Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16**, 1170–1178 (2013).
119. Koblinger, Á., Fiser, J. & Lengyel, M. Representations of uncertainty: where art thou? *Curr. Opin. Behav. Sci.* **38**, 150–162 (2021).
120. FitzGerald, T. H. B., Seymour, B., Bach, D. R. & Dolan, R. J. Differentiable neural substrates for learned and described value and risk. *Curr. Biol.* **20**, 1823–1829 (2010).
121. Huettel, S. A. Decisions under uncertainty: probabilistic context influences activation of prefrontal and parietal cortices. *J. Neurosci.* **25**, 3304–3311 (2005).
122. Monosov, I. E. Anterior cingulate is a source of valence-specific information about value and uncertainty. *Nat. Commun.* **8**, 134 (2017).
123. Acerbi, L., Vijayakumar, S. & Wolpert, D. M. On the origins of suboptimality in human probabilistic inference. *PLoS Comput Biol.* **10**, e1003661 (2014).
124. Yeon, J. & Rahnev, D. The suboptimality of perceptual decision making with multiple alternatives. *Nat. Commun.* **11**, 3857 (2020).
125. Dabney, W. et al. A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**, 671–675 (2020).
126. Haxby, J. V. et al. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
127. Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* **37**, 435–456 (2014).
128. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
129. Park, I. M., Meister, M. L. R., Huk, A. C. & Pillow, J. W. Encoding and decoding in parietal cortex during sensorimotor decision-making. *Nat. Neurosci.* **17**, 1395–1403 (2014).
130. Pillow, J. W. et al. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454**, 995–999 (2008).
131. Haynes, J. -D. A primer on pattern-based approaches to fMRI: principles, pitfalls and perspectives. *Neuron* **87**, 257–270 (2015).
132. Hubel, D. H. & Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **148**, 574–591 (1959).
133. DeWind, N. K., Adams, G. K., Platt, M. L. & Brannon, E. M. Modeling the approximate number system to quantify the contribution of visual stimulus features. *Cognition* **142**, 247–265 (2015).
134. Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C. & Fried, I. Invariant visual representation by single neurons in the human brain. *Nature* **435**, 1102–1107 (2005).
135. Baker, B., Lansdell, B. & Kording, K. A philosophical understanding of representation for neuroscience. Preprint at https://doi.org/10.48550/arXiv.2102.06592 (2021).
136. Nichols, M. J. & Newsome, W. T. Middle temporal visual area microstimulation influences veridical judgments of motion direction. *J. Neurosci.* **22**, 9530–9540 (2002).
137. Cortese, A., Amano, K., Koizumi, A., Kawato, M. & Lau, H. Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nat. Commun.* **7**, 13669 (2016).
138. Gherman, S. & Philiastides, M. G. Human VMPFC encodes early signatures of confidence in perceptual decisions. *eLife* **7**, e38293 (2018).

Author contributions

All authors contributed to the writing of this Review.

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/s41593-023-01444-y.

Correspondence and requests for materials should be addressed to Florent Meyniel.

Peer review information *Nature Neuroscience* thanks Máté Lengyel and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.