nature reviews neuroscience

https://doi.org/10.1038/s41583-026-01030-8

Perspective

Check for updates

Clarifying the conceptual dimensions of representation in neuroscience

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

Abstract

Despite the centrality of the notion of representation in neuroscience, the field lacks a unified framework for the concepts used to characterize representation, leading to disparate use of both terminology and the measures associated with it. To offer clarification, we propose a core set of conceptual dimensions that characterize representations in neuroscience. These dimensions describe relations between a neural response, features that may be represented and downstream effects of the neural response. A neural response may be shown to be sensitive or specific to a feature, invariant to other features or functional (it is used downstream in the brain). We use information-theoretic measures to illustrate these conceptual dimensions and explain how they relate to data analysis methods such as correlational analyses, decoding and encoding models, representational similarity analysis, and tests of statistical dependence or adaptation. We consider several canonical examples, including models of the representation of orientation, numerosity and spatial location, which illustrate how the evidence put forth in support or criticism of these models is systematized by our framework. By offering a unified conceptual framework to characterize representation in neuroscience, we hope to aid the comparison and integration of results across studies and research groups and to help to determine when evidence for a neural representation is strong.

Sections

[1]Department of Philosophy, New York University, New York, NY, USA. [2]Department of Neurobiology and Biophysics, Computational Neuroscience Center, University of Washington, Seattle, WA, USA. [3]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, USA. [4]Department of Philosophy, University of Pennsylvania, Philadelphia, PA, USA. [5]Center for Neural Science, New York University, New York, NY, USA. [6]Department of Psychological & Brain Sciences, Boston University, Boston, MA, USA. [7]Cognitive Neuroimaging Unit, INSERM, CEA, CNRS, Université Paris-Saclay, NeuroSpin center, Gif/Yvette, France. [8]Institute for Neuromodulation, GHU Paris psychiatrie et neuroscience, Sainte Anne Hospital, Université Paris Cité, Paris, France. [9]Department of Psychology, New York University, New York, NY, USA. [10]These authors jointly supervised this work: Florent Meyniel, Wei Ji Ma
e-mail: stephan.pohl@nyu.edu

Introduction

Humans and other animals perceive features of their environment, integrate sensory information with knowledge about rewards and goals, and produce behaviour to achieve goals in their environment. Neuroscience aims to explain how the brain brings about these abilities. In a widely used explanatory framework, sensory inputs are transformed into representations of features, while computations transform these representations and integrate them into cognitive processes, including reasoning, planning, learning and memory[1]. These processes result in a course of action, which is transformed into motor commands and executed as the individual's behaviour[2,3].

Neuroscientists agree that the notion of representation is a vital component in their explanatory practices[4,5]. Appeals to representations appear frequently. Marr[1] discusses the transformation of representations of local retinotopic properties to representations of shape properties used for object recognition in vision. Salzman et al.[6] find that stimulation with microelectrodes enhances the sensory representations of motion direction. Haxby et al.[7] suggest that the representations of faces and objects are distributed and overlapping. Kriegeskorte and Diedrichsen[8] discuss models of the representational geometry of neural responses. Yet, despite this widespread use of representational terminology, neuroscientists report high uncertainty about what kinds of findings count as evidence for representation[5]. The diverse set of tools used to characterize representations and the terminology to describe them leave room for clarification.

In this Perspective, we develop a framework that systematizes the types of finding neuroscientists point to in support or criticism of claims of the form that some neural response represents a feature of the environment. Findings can be categorized according to the kind of relation they establish between a neural response, features that may be represented, and downstream computations integrating that neural response.

We identify four kinds of relations, which we call the conceptual dimensions of representation. These dimensions are tools for understanding and describing how different methodological approaches contribute to a common project of establishing claims about representations. Experimental designs and data analysis methods can be categorized by the dimensions they provide evidence for. By distinguishing four dimensions relevant to representation, we describe a space in which findings can be located. This is not to say that any region in this conceptual space corresponds to stronger evidence than other regions, but it is helpful to be explicit about whether two bodies of work address the same aspects of representation. For ease of exposition, we focus on representations that carry information about the state of the world, whether past, present or future. Not all representations do. Imaginings, hypothetical reasoning and representations of goals can occur independently of the state of the world.

The conceptual dimensions of representation that we propose are sensitivity, specificity, invariance and functionality. Take the claim that the fusiform face area (FFA) represents faces[9]. Findings in support of a claim that a given neural response (FFA neural activity) represents a given feature (faces) may show that, first, the neural response is 'sensitive' to the feature, that is, it carries information about the feature. Second, they may show that the neural response is 'specific' to the feature, that is, most variations of the neural response occur only when the feature is changed. Third, findings may reveal that the neural response to the feature of interest is 'invariant' to other features; a neural response representing a face, for instance, responds to faces irrespective of low-level visual features such as position or lighting. Last, they may show that the neural response is 'functional', that is, it makes information about the feature available for integration with other cognitive processes, for instance, the transformation into motor commands in the form of a behavioural response. The representation of faces in FFA is a classic example of representation in neuroscience. Early papers that developed methodology to identify populations of neurons on the basis of their role in information processing[9] and to analyse distributed representations[7] studied the FFA. Sensitivity, specificity, invariance and functionality have all been addressed in the literature on representations of faces in FFA.

We introduce an information-theoretic formalization of these conceptual dimensions of representation, not as a new approach but rather as a conceptual tool to make explicit dimensions, systematizing what neuroscientists have been doing all along in studying representations. We develop a framework that researchers can use to integrate results across studies and data analysis techniques. The information-theoretic formalism we use builds on a long line of previous work. Although originally developed in the context of engineering artificial information-processing systems[10], information theory has also been used to analyse natural information-processing systems in neuroscience[3] and philosophy[11,12]. It has been used in neuroscience to characterize the information captured by decoding models[13], the efficiency with which spike trains represent features[14–17], and the information transferred within the brain[18–21], and it has been proposed as a general statistical framework for neural data analysis[22–25]. Here, we do not endorse particular statistical techniques but instead offer conceptual clarification of research questions about representations to unify common operationalizations in terms of decoding or reconstruction models[26,27], brain signatures[28], representational similarity[8,29], or information-theoretic data analyses[19,30] that are used to address these research questions.

In this Perspective, we begin by introducing sensitivity, specificity, invariance and functionality in information-theoretic and causal terms. We then consider how common methodologies are used to evaluate the conceptual dimensions of representation, thereby explaining the integration of these methodologies into one unified project. We next illustrate how tests of the dimensions combine to form comprehensive bodies of evidence by looking at several canonical examples of representations: orientation in primary visual cortex (V1), numerosity in the parietal cortex and spatial location in the hippocampus. Finally, we briefly consider limitations and potential extensions of our framework of the conceptual dimensions of representation.

The conceptual dimensions of representation

We introduce our framework using a toy example: we are interested in how the ripeness of an apple is represented in the brain. In our experimental setup, apples are presented to a participant, brain activity is recorded and the participant chooses whether to eat the apple (Fig. 1). All the ways in which our experimental environment could be different, including the stimulus presented to the participant (apples), can be summarized in a feature space in which each dimension of that space is one feature. The feature of interest (s) – in our experiment, the ripeness of an apple – is distinguished from any other feature (n) by being the investigation's target feature. We are testing the hypothesis that the neural response (r) represents s. The behavioural response (b) is the participant's decision to either eat or not to eat the apple (Fig. 1).

In typical experimental contexts, s could be any feature of interest, such as the orientation of a grating, the category of an object or a more abstract feature such as the probability with which another

feature changes its value. Meanwhile, *n* could be any other feature included in the analysis, such as the contrast of an image, the perspective from which an object is viewed, or the modality of a stimulus or the task context. Often, multiple features are included, so that *s* and *n* are multidimensional. In our toy example, we identify three features, n_1, n_2 and n_3, which differ in how they are causally and statistically related to *s* and *r* (Fig. 1). Going forward, we collapse these into one multidimensional variable *n*. Typically, *r* is a high-dimensional vector reflecting the activity pattern of a population of neurons, for example, measured with functional MRI (fMRI), electroencephalography (EEG) or implanted electrodes. Finally, *b* could be any behavioural response to *s*, such as an eye movement or a button press, by which the participant, for instance, reports the value of *s* or makes some decision based on *s*.

Some methodological approaches are framed not in terms of the representation of features but in terms of representations that are selective for a target category. A neural event may indicate the presence of a member of that category, such as the firing of neurons in the FFA indicating the presence of faces[9]. Note that a population of neurons that is selective for a target category is thereby also sensitive and specific to the binary categorical feature whose values are the instantiation or non-instantiation of the target category. Modelling the representation as a representation of features can also be extended to many-valued categorical features, such as object category, or to features with a graded scale, such as orientation. For many-valued categorical features, it does not make sense to speak of the presence or absence of a target category, which is why we prefer the more general framing in terms of features (Fig. 1).

Often, experiments start with *s* and search for that *r* in terms of which *s* is represented (for instance, to investigate where in the brain faces are represented[9]). By contrast, some start with a given *r* and try to find which *s* is represented (for instance, to investigate the receptive field of V1 neurons[31] or by using model-based approaches that fit models of *s* to *r*[32]). Both approaches can be combined such that models of neural subpopulations (*r*) and feature spaces (*s*, *n*) are advanced in unison[33].

The evaluation of the conceptual dimensions of representation — sensitivity, specificity, invariance and functionality — depends on which features are included in the analysis, and so it may be misleading when relevant features are excluded. In our toy example, we are looking for a representation of ripeness; say we do not consider colour as an additional feature and test our hypothesis with a species of apple in which red apples tend to be ripe but green apples are not. By excluding a relevant feature (colour) from our analysis, we may identify some *r* that is sensitive to ripeness (*s*) even though it is a representation of colour (n_2; Fig. 1). A better experimental setup would modulate ripeness and colour independently by using a species of apple that may be green even when ripe. With colour included as a relevant feature, *r* no longer appears to be sensitive to ripeness (when holding fixed the colour, *r* no longer carries information about ripeness) and fails to be invariant to colour (even while holding ripeness fixed, *r* continues to carry information about colour). In summary, whether measures of sensitivity and invariance reveal *r* to be a representation of colour or mistake it for a representation of ripeness depends on whether all relevant features are included in the analysis.

We first introduce the conceptual dimensions of sensitivity, specificity and invariance in information-theoretic terms (Fig. 2; see Box 1 for a summary of basic notions of information theory). These dimensions describe the relation between *s*, *n* and *r*. Functionality describes the causal role of *r* in downstream cognitive processes and the production of *b*, and is introduced subsequently.

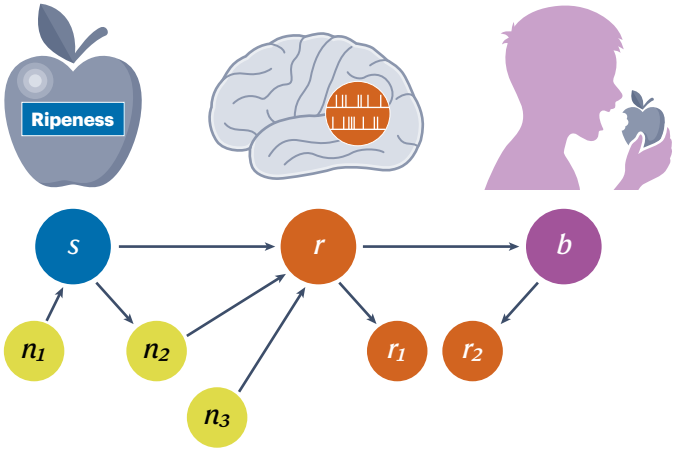

Fig. 1 | Generative model of our toy example. *s* is the ripeness of an apple, *r* is a neural response that might be a representation of *s*, and *b* is a behavioural response; here, our participant eats or does not eat the apple depending on whether they perceive it to be ripe. To show that *r* is a representation of *s*, alternative hypotheses about other features (*n*) are considered that stand in diverse statistical relations to *s* and *r*. In our toy example, *n* is exemplified by the age of an apple (n_1), the colour of an apple (n_2) or the colour of the packaging in which the apple is presented to the participant (n_3). In addition, *r* is also shown to be used as a representation of *s*. A brain area into which activity from *r* spills over (r_1) but which has no role in *s*-related processing, or areas involved in monitoring the participant's own behaviour (r_2), may be related to *s* but fail to be functional as a representation of *s*. In the figure, nodes represent variables and arrows direct causal dependencies, which determine statistical dependencies between variables.

The order in which we introduce these conceptual dimensions of representation is chosen for didactic reasons. In practice, there is no general order in which the dimensions are evaluated, with findings about one dimension often inspiring tests of others in a complex back-and-forth.

Relations between features and the neural response

The relation between features and the neural response is described in terms of the sensitivity of *r* to *s*, its specificity to *s*, and its invariance to *n*. Sensitivity and specificity may further be evaluated conditional on *n*.

Sensitivity. *r* is sensitive to *s* given that *r* is informative about *s*. In our toy example (Fig. 1), *r* is sensitive to the ripeness of an apple given that ripeness can be inferred from *r* (for example, because the neural response rate is higher for ripe apples than for non-ripe apples). An empirical finding of sensitivity is the observation that neurons in the middle temporal area (MT) respond preferentially to directional motion, allowing the feature of motion direction to be predicted from their responses[34].

For *r* to be sensitive to *s*, there must be mutual information (*I*) between *r* and *s* (Box 1). However, the absolute amount of mutual information is not very useful. What matters for sensitivity is that the mutual information between *r* and *s* captures a large proportion of the total information about *s*. If *r* captures the total information about *s* (such that $I(s;r) = H(s)$ in Eq. 1), it is possible to infer the value of *s* from *r* without error (Fig. 3a, right). This upper bound on the measure of sensitivity is achieved by normalizing the mutual information between *s* and *r* by

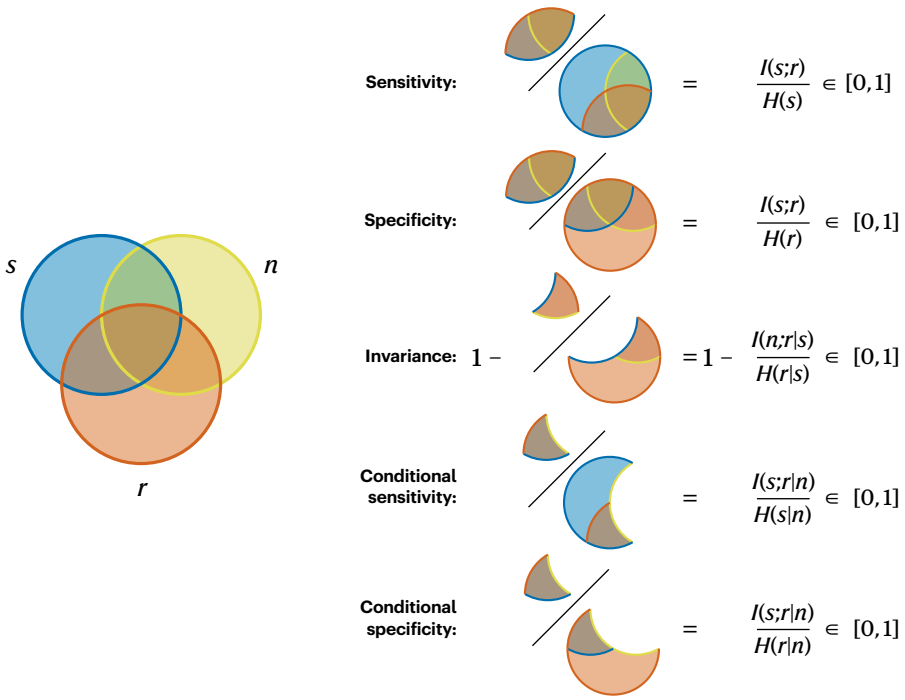

Fig. 2 | The notions of sensitivity, specificity and invariance in information-theoretic terms. The variables *s*, *n* and *r* correspond to the feature of interest, other features and the neural response, respectively. Circles in the Venn diagram represent the total variability associated with a given variable measured in terms of the entropy (*H*) of that variable (left). Overlapping areas represent the mutual information (*I*) between variables. Information-theoretic quantifications of sensitivity, specificity, invariance and the conditional forms of sensitivity and specificity are given by the equations (right). The shapes depict the same quantities in terms of the respective areas in the Venn diagram (middle). The numeric values of sensitivity, specificity and invariance range from 0 to 1, with 1 indicating maximal satisfaction and 0 indicating no satisfaction.

the entropy (*H*) of *s* (Box 1 and Fig. 2). In statistics, this normalization of mutual information is sometimes called the uncertainty coefficient[35].

$$\text{Sensitivity}: \frac{I(s\,;\,r)}{H(s)} \tag{1}$$

That *r* is highly sensitive to *s* therefore means that the information *r* carries about *s* captures a large proportion of the variability of *s*. Because of the normalization, sensitivity takes values between 0 (no sensitivity) and 1 (perfect sensitivity) but perfect sensitivity cannot generally be expected, as representations may be inaccurate. For instance, people inexperienced in telling species of trees apart may represent a tree as being a beech; however, that representation would often be inaccurate (and its sensitivity to the feature of tree species substantially below 1). Despite this, tree species might still be the feature to which that representation is most sensitive.

Specificity. *r* is specific to *s* if many of the changes in *r* are explained by changes in *s*. In other words, *s* explains a large proportion of the variability of *r*, or, given *s*, *r* remains mostly constant (Fig. 3b, right). Recall that *s* can be multidimensional (in our toy example, we might add a second feature of interest besides ripeness, such as sweetness; Fig. 1). A representation that is sensitive to multiple features — displaying mixed selectivity[36,37] — will be specific only to the combination of these features.

In information-theoretic terms, specificity is quantified by the mutual information between *s* and *r*, normalized by the entropy of *r* (Fig. 2).

$$\text{Specificity}: \frac{I(s\,;\,r)}{H(r)} \tag{2}$$

That *r* is highly specific to *s* therefore means that a large proportion of the variability of *r* carries information about *s*. An equivalent expression of specificity is as follows:

$$1-\frac{H(r|s)}{H(r)} \tag{3}$$

That is, specificity can also be thought of as the proportion of the variability of *r* that is explained by *s*. A lack of specificity suggests that some factors other than *s*, be they noise, internal processes or *n*, drive the remaining variability of *r* (in Fig. 2, $H(r|s)$, the area of $H(r)$ that does not overlap with $H(s)$, would be large). Specificity quantifies only the dependence of *r* on *s* but not the dependence on other factors. A finding of low specificity (Fig. 3b, left) may be followed by tests aiming to identify the remaining factors driving the variability of *r*.

Specificity has also been used under the label 'coding efficiency'[15,17]. This work evaluated what proportion of the variability of spike trains carries information about the stimulus; the larger the proportion of the variability of spike trains that carries information about the stimulus, the more efficient the neural code is — the mapping from *r* to estimates of *s* (Supplementary Box 1). In a neural code of low efficiency, much of the variability of *r* fails to carry information about *s*, which may be metabolically costly[38].

Sensitivity and specificity both depend on the mutual information between *r* and *s*; they differ only in the normalizing factor (Fig. 2). High sensitivity means that *s* can be inferred from *r* with high accuracy, whereas high specificity means that *r* can be inferred from *s* with high accuracy (Fig. 3a,b). Often, sensitivity is approximated by the performance of decoding models and specificity is approximated by the performance of encoding models[39]. The close connection between sensitivity and specificity, but also the importance of distinguishing these two relations, has been observed previously. Poldrack[40,41] distinguishes between forward inferences (based on encoding models) and reverse inferences (based on decoding models) and points out how the validity of a reverse inference (decoding a representation from *r*) depends on the specificity of *r*. Averbeck et al.[42] point out how noise correlations may have different impacts on how information

about features is encoded in *r* (reflecting specificity) or decoded from *r* (reflecting sensitivity).

The notions of sensitivity and specificity in our framework should not be confused with sensitivity and specificity in a binary decision or test context. In those contexts, 'sensitivity' refers to the true positive rate (the proportion of positive instances truly detected as positive) and 'specificity' refers to the true negative rate (the proportion of negative instances truly detected as negative). A high-level analogy exists between the two notions; in both, sensitivity indicates that a test or representation is responsive to its target whereas specificity indicates that it is not responsive to other things. Yet, mathematically, these notions are quite different.

A perfect specificity of 1 is not usually expected to be found because of factors unrelated to *s* that affect the neural response, including neural noise[43] and internal processing reflecting brain states such as arousal, motivation or mind wandering.

Invariance. The representation of *s* by *r* is invariant to *n* if *r* does not depend on *n* for a given value of *s* (if *r* is not sensitive to *n*, conditional on *s*). In our toy example (Fig. 1), to show that the representation of ripeness by *r* is invariant to the apple's colour, one would have to show that *r* indicates ripeness independently of colour. A typical empirical finding of invariance is that neural responses in the primate inferior temporal cortex, which are sensitive to object category, did not respond to changes in the orientation of the object[44,45], or that neural responses, which are sensitive to surface colour, remained unchanged across variations in the ambient light colour (displaying colour constancy)[46].

When *s* and *n* are statistically related and *r* is sensitive to *s*, it may also be sensitive to *n* in virtue of the statistical dependence between *s* and *n*. Thus, *r* cannot generally be expected to lack sensitivity to *n* even though it may be invariant to *n*. To demonstrate invariance, one must show that *r*'s dependence on *n* is accounted for by *r*'s dependence on *s*. In other words, *r* is invariant to *n*, if, conditional on *s*, *n* does not account for any of the variability of *r* (Fig. 3c, right).

In information-theoretic terms (Fig. 2), invariance is quantified as follows:

$$\text{Invariance}: 1 - \frac{I(n\,;r|s)}{H(r|s)}, \tag{4}$$

which can, equivalently, be expressed as follows:

$$\frac{H(r|n,s)}{H(r|s)} \tag{5}$$

An invariance of 1 means that no more of the variability of *r* can be accounted for by the combination of *s* and *n* than by *s* alone. An invariance below 1 means that some of the variability of *r* that is unrelated to *s* can be explained by *n*.

Two applications of invariance are of particular interest, namely those in which *n* stands for the modality of a stimulus or for the task. A representation of location might be invariant to whether the location of an event is presented to a participant as a visual or an auditory

Box 1 | Information theory

Here, we provide a brief introduction to central concepts of information theory (see ref. 50 for a more extensive treatment).

In the introduction of our framework, we assume that variables are discrete (see 'Our framework applied across methodologies' for discussion of the application of the framework in data analysis, which includes continuous variables). We write random variables as lower-case letters (for example, *x*). Values of the variables are written with an index (for example, x_i). x_i also abbreviates the event $x=x_i$.

Entropy

Entropy (*H*) measures the uncertainty about a random variable, which can also be understood as the amount of variability or randomness associated with a variable. Entropy is high when a variable has many values with equally high probabilities (*p*).

Definition: $H(x) = -\sum_i p(x_i)\log p(x_i)$

Conditional entropy

The conditional entropy measures how much uncertainty there is on average about some variable conditional on another variable — that is, it measures the average uncertainty about the first variable that remains after the value of the second variable is specified.

Definition: $H(x|y) = -\sum_i p(y_i) \sum_j p(x_j|y_i)\log p(x_j|y_i)$

Mutual information

The mutual information (*I*) between two variables measures how much uncertainty about one variable is reduced by conditioning on the other variable. That is, mutual information is high when there is much less variability of the values of one variable conditional on the other variable than there is variability not conditional on any variable. Mutual information is symmetric ($I(x;y)=I(y;x)$).

Definition: $I(x\,;y) = H(x) - H(x|y) = H(y) - H(y|x)$

Mutual information can also be understood as a measure of the statistical dependence between two variables. When *x* and *y* are statistically independent, then $I(x;y)=0$. When there is some statistical dependence between *x* and *y*, then $I(x;y)>0$.

Joint entropy

The entropy of a joint distribution is calculated as follows.

Definition: $H(x, y) = -\sum_{i,j} p(x_i, y_j)\log p(x_i, y_j)$

Conditional mutual information

Conditional mutual information measures the mutual information between two variables conditional on a third variable. That is, it measures how much, on average, uncertainty about a first variable is reduced by conditioning on a second variable, given that the value of the third variable is already specified.

Definition: $I(x\,;y|z) = H(x|z) - H(x|y,z) = H(x,z) + H(y,z) - H(x,y,z) - H(z)$

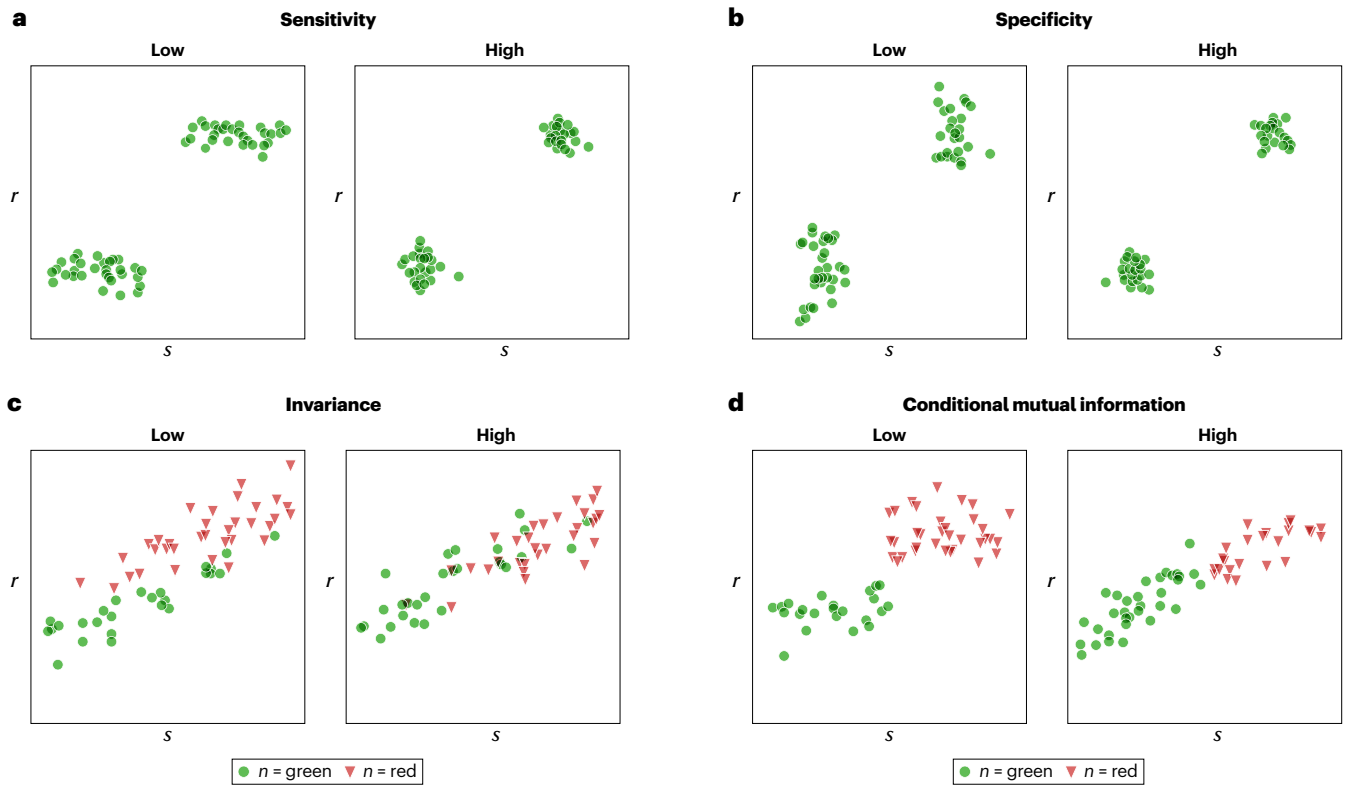

Fig. 3 | The conceptual dimensions of representation sensitivity, specificity and invariance capture relations between features and the neural response. **a**, Sensitivity is high when *s* can be inferred from *r* with high accuracy. When sensitivity is low, the distribution of *s* given *r* is more spread out, such that a decoding model of *s* given *r* would be less accurate (left). By contrast, given *r*, little variability of *s* remains when sensitivity is high, such that a decoding model of *s* given *r* can be highly accurate (right). Changes like this in the stimulus distribution across low- versus high-sensitivity conditions may be less typical for controlled experiments but are to be expected in more natural environments. **b**, Specificity is high when *s* explains a large proportion of the variability of *r*. When specificity is low, conditional on *s*, the remaining amount of variability of *r* is large (left), whereas for a given value of *s*, there is little difference between values of *r* when specificity is high (right). In part **a**, only sensitivity changes but specificity remains unchanged, whereas in part **b**, specificity changes and sensitivity remains unchanged; in the low conditions, additional variability is introduced, which spreads out the distribution either along *s* (sensitivity) or *r* (specificity) but leaves the distribution along the other variable unchanged. Such a clear dissociation of sensitivity and specificity is rare in practice and exaggerated here in our simplified example. Yet, sensitivity and specificity do commonly come apart, for instance, when neural noise or other drivers of neural variability undermine specificity, whereas sensitivity remains unaffected because the neural code for *s* is robust to the interfering drivers of variability (such as by averaging out neural noise). **c**, Invariance is high when, conditional on *s*, there is no dependence of *r* on *n*. There may be a dependence of *r* on *n*, even though invariance is high, if this dependence disappears conditional on *s* (for a given value of *s*, it is no longer the case that *r* tends to be larger for *n* = red than for *n* = green; right). When even conditional on *s*, *r* depends on *n* (*r* tends to be larger for *n* = red than for *n* = green; left), the invariance is low. **d**, The mutual information between *s* and *r* conditional on *n* is high when, even conditional on *n*, there is a strong dependence between *s* and *r*. When conditional on *n*, the dependence between *r* and *s* is broken (for a given value of *n*, *s* and *r* are statistically independent; left) and the mutual information between *r* and *s* conditional on *n* is low. When the dependence persists (even conditional on *n*, *r* is informative about *s*; right), the conditional mutual information is high.

stimulus (*r* is invariant to *n* when *r* represents *s* both in the condition when *n* = auditory and when *n* = visual). Invariance to modality is great evidence for the claim that *r* is a representation of *s*. However, lack of invariance to modality is more difficult to interpret — it might be that the same features are independently represented in the perceptual systems that are specialized for different modalities. Another application is task invariance; here, *r* is task-invariant when it is sensitive to *s* irrespective of *n*. A lack of task invariance does not need to indicate that *r* is a poor candidate for a representation of *s*; it might be that the brain represents *s* only when it has to. That *r* is particularly sensitive or specific only in conditions where it is task relevant may even suggest that it is used in the response to the task (that it is functional). For instance, in the study of feature-based attention, task context has been found to modulate the sensitivity of representations to particular dimensions of the feature space such as motion or colour[47].

Both the invariance and specificity of *r* for complex features (such as object category[48] or numerosity[49]) tend to increase along perceptual processing hierarchies. By the data processing inequality[50] — which states that, along a causal chain, mutual information can only be maintained or decreased but not increased — sensitivity can only stay the same or decrease from early to later processing stages. However, neural responses at later processing stages tend to be more specific to complex features and invariant to low-level features (such as retinotopic maps of orientation, contrast and so on) as representations of the complex features are disentangled from the low-level features[48]. For instance, object category can already be decoded

from early visual areas using sophisticated models; however, in the inferotemporal cortex, populations of neurons are highly specific to object category and invariant to low-level features such as the object's orientation. Relatedly, invariance may often be interpreted as the result of a computation or inference that reconstructs the value of latent features (such as object category) from intermediate features more directly observable by the participant (such as the retinal image)[51–55].

Conditional sensitivity and specificity. Sensitivity and specificity can be evaluated conditional on n. A common goal for testing conditional sensitivity and conditional specificity is to rule out that r depends on s merely because r depends on n and s happens to also depend on n.

In our toy example (Fig. 1), we may test whether even just for green apples, r continues to depend on the ripeness of the apple (whether r differs between green apples that are ripe and green apples that are not ripe). If r were merely a representation of colour, conditional on colour, it would not carry any information about ripeness (Fig. 3d). A typical empirical test of conditional sensitivity, for instance, may show that a neural response to the numerosity of a dot cloud carries information about numerosity even while holding fixed related quantities such as density, shape or surface area covered by the dot cloud[56].

As with their unconditional analogues (Fig. 2), conditional sensitivity and conditional specificity are quantified by different normalizations of the mutual information between s and r but conditional on n:

$$\text{Conditional sensitivity}: \frac{I(s;r|n)}{H(s|n)} \tag{6}$$

$$\text{Conditional specificity}: \frac{I(s;r|n)}{H(r|n)} \tag{7}$$

That r is sensitive to s, conditional on n, means that the information r carries about s, after conditioning on n, captures a large proportion of the variability of s. That r is specific to s, conditional on n, means that, after conditioning on n, a large proportion of the variability of r carries information about s. The conditional variants of sensitivity and specificity, as well as invariance, are of particular relevance when r depends on interactions of s and n (Supplementary Box 2).

In practice, sensitivity and specificity are always evaluated conditional on the experimental setup. This may undermine the ecological validity of the findings. For instance, it introduces the risk of finding dependencies conditional on some aspect of the experimental setup that may not generalize (for example, because features covary in the experiment but not in the natural world). Or it may obscure dependencies that are present in natural environments but may be absent in constrained experimental setups. Controlling the experimental environment, varying experimental conditions and randomizing them can mitigate these risks; however, they cannot be avoided completely.

Functionality

If r is a representation of s, we not only expect it to carry information about s but also expect that this information is used in the brain[3,4,26,57,58] (that r is functional). Generally, when r is a representation of s, subsequent processes that rely on information about s receive that information from r. In our toy example (Fig. 1), r is a representation of ripeness when the participant makes the decision to eat an apple (b) based on whether r represents the apple to be ripe or not (r is functional in virtue of being used in this decision).

Use implies causality. Therefore, functionality has to be evaluated in causal terms[19,59,60]. Causal claims can be established with experimental interventions[61,62] — to show that r is functional, one can show that interventions on r modulate what information about s is available to downstream processes (such as a behavioural report of s). For example, Salzman et al.[6] showed that stimulation of neurons with a specific preferred motion direction biases the behavioural report of the perceived motion direction towards that preferred motion direction.

Information-theoretic quantities, much like correlations, are not sufficient to capture causality and, thereby, functionality, but they are still useful. General statistical methods for inferring causal dependencies from observational data have been discussed in the context of cognitive neuroscience[63–67]. A precursor to claims about functionality are claims about functional connectivity (claims about statistical relations between populations of neurons that suggest the transfer of activation and information between them)[68–71]. More specific statistical methods have been developed to trace the transfer of information about a feature between populations of neurons[18,72] or to trace the transfer of information from r to b[73–75] (for example, in terms of the probability of correctly predicting b from r[76]. Many methods consider the relations between s and r or between b and r in isolation. To address functionality more specifically, it has been proposed to analyse the intersection information[19,20], which quantifies not only the information transfer from s to r, or from r to b, but quantifies how much information is in common between s, r and b.

The impact on b of most representations is mediated by other representations. For instance, sensory representations may be integrated with representations of goals to form plans for actions and to produce motor commands, which are turned into b. A full understanding of functionality requires tracing the flow of information about s from sensory input to motor commands and b (in the form of multistage process models). For example, Wilming et al.[77] showed, in a perceptual decision-making task, how sensory signals are represented in early visual areas, accumulated and transformed into action plans in parietal and motor areas, and, ultimately, turned into a behavioural response by pressing a button with the participant's left or right hand. Further, choice-reflective motor activity fed back into early visual areas, modulating how much sensory representations are integrated into the decision-making process.

Statistical methods for inferring causal connections are fundamentally limited. A central challenge is that a statistical dependence between two variables need not be due to a causal connection between these variables but may be due to a causal dependence on a third variable. For instance, in our toy example (Fig. 1), there is a statistical dependence between r_1 and b — they may even share information about s — not because r_1 causes b but because both share s (and, in this case, also r) as common causes. Thus, r_1 may carry information about both s and b without being functional. Some tests of functionality try to exclude a shared statistical dependence on s by controlling s. For example, Bradley et al.[78] showed that the reported motion direction of bistable stimuli — 2D projections of rotating transparent cylinders where the perceived motion direction changes spontaneously between the two possible directions of rotation around the cylinder's central axis — could be predicted from neural activity in MT. Similarly, functionality can be evaluated by showing that errors (variability of b that is unrelated to s) can be predicted from r[74].

No observational data can exclude the possibility that some unobserved r explains a statistical dependence between variables without a direct causal connection. Where possible, experiments in which

an intervention is performed directly on *r*, leaving stimulus and task context unchanged, are particularly valuable. Changes in *b* following such an intervention are strong evidence of a causal role of *r* in bringing about *b* (evidence of functionality). For example, in non-human primates, stimulation of neurons in MT modulated the perceived direction of motion, providing evidence that MT is functional as a representation of motion direction[6], whereas stimulations in the lateral intraparietal area modulated how quickly a participant reported a specific direction of motion, providing evidence that the lateral intraparietal area is functional as a representation of the strength of sensory evidence about motion direction[79]. Similarly, chemically induced lesions of MT in monkeys[80] and suppression of area V5 with transcranial magnetic stimulation in humans[81] both impaired motion perception specifically. Interventions performed directly on *r* may not address effects on *b* but may instead address effects on subsequent neural processing; for example, stimulations of neurons in the frontal eye field increased the information about stimulus features carried by retinotopically corresponding neurons in area V4 of non-human primates[82].

We used information theory as a conceptual tool to introduce our framework. The conceptual dimensions of sensitivity, specificity and invariance describe the relation between *s*, *n* and *r*. Functionality describes the use of *r* in downstream cognitive processes and the production of *b*. Together, these conceptual dimensions can be used to describe the typical empirical findings based on which claims about representation are made in cognitive neuroscience.

Our framework applied across methodologies

We turn to a survey of common data analysis methods. We consider how analyses of correlation, decoding and encoding models, tests of statistical dependence, representational similarity analyses, and findings about adaptation are used to provide evidence about representations and how this is systematized by our framework. We highlight formal connections between the information-theoretic formalism of our framework (Fig. 2) and these methods to explain how they are related conceptually, but we do not work out in general under what conditions these methods would be mathematically equivalent. This discussion explains how our framework applies to a broad range of commonly applied methods; however, it is only a starting point of an analysis of the full range of methods applied in cognitive neuroscience.

Information theory in neural data analysis

Information theory can be applied directly in neural data analysis. This may be motivated by not only thinking of the brain as an information processing system but also the utility of information theory for evaluating statistical relations between variables without assumptions about the structure of those relations (its model independence)[3,22–25,83,84]. For instance, information theory has been used to characterize the manner in which spike trains carry information about the stimulus[15,17], the effects of attention on population codes[85,86] and object representations[87], the relation between oscillations in EEG data and face expression perception[88], and the relation between EEG and fMRI data[89] or to characterize the processing of rhythmic components of speech[90]. Where entropies of variables and their mutual information are estimated, this affords the most straightforward interpretation in terms of our framework (Fig. 2).

As estimates of entropy and mutual information generally require estimating full probability distributions (Box 1) rather than moments of the distribution like variance, they are particularly challenging for high-dimensional and continuous variables from limited sample sizes. Often, to obtain estimates of information-theoretic quantities, data are discretized into relatively few bins, such that there are sufficient data for each bin. However, such binning introduces biases. Methodologies are being developed within neuroscience and in the larger context of statistics and machine learning with an eye towards benchmarking the reliability of estimators with respect to these challenges[91–94] (see 'Decoding models' for a discussion as many such methods are based on them).

A particular challenge for continuous variables is the normalization of information-theoretic quantities. Take specificity, which quantifies how much of the total variability of *r* carries information about *s* (Fig. 2). In the discrete case, the total variability of a variable can be measured by its entropy. In the continuous case, no trivial measure of total variability is applicable to probability distributions in general (but see Nagel et al.[95] for a discussion of general estimators for normalized mutual information).

The rate at which *r* provides information to downstream processes may be used to normalize information-theoretic quantities (Supplementary Box 3). This rate is limited, for instance, by noise that may manifest in the temporal resolution with which downstream processes depend on *r*. Rieke et al.[17] evaluated the mutual information between *s* and *r* ($I(s;r)$) normalized by the entropy of *r* ($H(r)$) relative to a discretization of *r* (with the size of the discrete bins of *r* assumed to correspond to the lower limit of the temporal resolution of the neural code). While the normalization by $H(r)$ relies on a discretization of *r* (potentially introducing biases), it does reflect not only stimulus-dependent variability but also endogenously driven processing and noise corrupting the signal. By contrast, Tafazoli et al.[96] normalized $I(s;r)$ by the total information *r* carries about the stimulus ($I(s,n;r)$). The normalization by $I(s,n;r)$ allows the use of a broader range of estimators for information-theoretic quantities that may avoid the biases of discretization and captures only the proportion of stimulus-dependent variability driven by *s* or *n* but not noise or other stimulus-independent drivers of variability.

Linear correlation

A very common quantitative measure of the relationship between two variables, whether discrete or continuous, is their Pearson correlation. High correlation between *s* and *r* means that *s* can be predicted well from *r* and that *r* can be predicted well from *s* (both sensitivity and specificity are high). As correlation is symmetric, it does not distinguish between sensitivity and specificity, which are asymmetric (when *r* is highly sensitive to *s*, it need not also be highly specific to *s*; Fig. 3).

Lower degrees of correlation are more difficult to interpret because mutual information may be high even though correlation is low when the dependence of *r* on *s* is non-linear[97] (for example, U-shaped; Fig. 4a). If the dependence between *r* and *s* is linear with added Gaussian noise, then the mutual information is monotonically related to the Pearson correlation coefficient (ρ)[50]. In this scenario, low degrees of correlation also imply low mutual information and, therefore, low sensitivity and specificity.

Correlation can also be used to evaluate invariance, where a low degree of correlation of *r* with *n* indicates invariance to *n* (but, again, *r* may in fact depend non-linearly on *n*). The correlation between *r* and *b* is an estimate of functionality – but only a crude one, as it poorly reflects causal dependence.

Decoding models

The most common measures of sensitivity are measures of the performance of a decoding model, that is, a model that infers *s* from *r*.

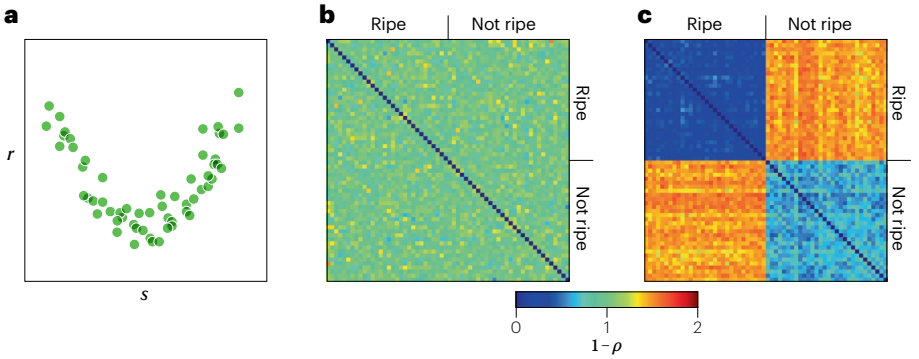

Fig. 4 | How common measures are related to our framework. a, A non-linear relationship between *s* and *r* (*r* depends on *s* by a quadratic function). There is high mutual information between *s* and *r*, but *s* and *r* are not correlated. **b**,**c**, Interpretation of representational dissimilarity matrices (RDMs). In this hypothetical example, 30 ripe apples and 30 apples that are not ripe were presented to a participant. *r* was recorded from different brain areas, and the correlation distance ($1-\rho$) of *r* for all pairs of stimuli (all pairings of apples, across all values of ripeness) was computed to construct RDMs. There is no difference between dissimilarities for ripe and not ripe apples in the RDM for the brain area in part **b**; *r* is not sensitive to the category distinction. By contrast, in the RDM for the brain area in part **c**, there is high dissimilarity across the category boundary of ripe versus not ripe; *r* is sensitive to this distinction. Yet, there is little dissimilarity between exemplars within categories; hence, *r* appears to be invariant to within-category distinctions.

Such models have been used to decode faces[98] and object identity[99] from the monkey temporal cortex, spatial location from the human hippocampus[100], auditory location from the monkey auditory cortex[101], and object category from early and late visual areas in humans[102].

The accuracy of a decoding model is a good measure of sensitivity because it has an easy-to-interpret lower and upper bound. When the decoding accuracy is at the chance level, this means that the given decoding model cannot extract any information about *s* from *r*, suggesting that the sensitivity is 0. The higher the decoding accuracy is above the chance level, the higher the sensitivity, with 100% decoding accuracy meaning one can perfectly infer *s* from *r* (the sensitivity of *r* to *s* is 1).

A general measure of sensitivity can be derived from the expected logarithmic loss of a decoding model estimated by the cross-validated loss on a separate test dataset, rather than the loss on training data. Consider a decoding model $p_\theta(s|r)$ with parameters θ in which $\mathrm{E}[-\log p_\theta(s|r)]$ is its expected logarithmic loss under the true joint probability distribution $p(s,r)$. If the decoding model is fitted well and $p_\theta(s|r)$ is identical to $p(s|r)$, then $\mathrm{E}[-\log p_\theta(s|r)] = H(s|r)$ (that is, the expected logarithmic loss of the decoding model is an estimate of the entropy of *s* conditional on *r* ($H(s|r)$)). An expected logarithmic loss of 0 means that the sensitivity of *r* to *s* is 1; *s* can perfectly be decoded from *r*. That is, the lower bound on the expected logarithmic loss of a decoding model corresponds to the upper bound on sensitivity. The lower bound on sensitivity corresponds to the chance level of a decoding model — when the sensitivity is 0, the expected logarithmic loss is equal to the loss of a model based on the prior probability distribution of *s* ($p(s)$) alone ($\mathrm{E}[-\log p(s)]$), which estimates $H(s)$. Sensitivity can be estimated from the expected logarithmic loss of a decoding model as follows:

$$\text{Sensitivity}: \frac{I(s\,;\,r)}{H(s)} = \frac{H(s) - H(s|r)}{H(s)} = \frac{\mathrm{E}[-\log p(s)] - \mathrm{E}[-\log p_\theta(s|r)]}{\mathrm{E}[-\log p(s)]} \quad (8)$$

Any measure of a decoding model's performance that covaries with the expected logarithmic loss can be used as a measure of sensitivity in an analogous manner. Expected logarithmic loss and accuracy of a decoding model are useful particularly for classification problems (decoding of discrete variables). The mean squared error or absolute error are common measures in the case of continuous variables, where decoding models commonly only offer a point estimate for *s*, rather than an estimate of the full probability distribution of *s* given *r* ($p(s|r)$).

By contrast, when $p_\theta(s|r)$ and $p(s|r)$ mismatch, then the expectation of the logarithmic loss of a decoding model merely offers an upper bound for the true entropy of *s* conditional on *r* ($H(s|r)$). That is, the estimate of sensitivity with a decoding model is merely a lower bound on the true sensitivity. A decoding model $p_\theta(s|r)$ with parameters θ may fail to closely approximate the true probability of *s* given *r* ($p(s|r)$) for two main reasons. First, the available dataset (both training and test data) may mismatch the underlying distribution because it is too small or biased. Second, the decoding model might not be sufficiently complex to fit the true relationship between *s* and *r* (for example, a linear model will not pick up on non-linear relationships). A decoding model may also suffer from excessive complexity, especially with small datasets. Bayesian decoding includes explicit measures of model complexity (to avoid overfitting) and of how well a decoding model fits the data[103].

As the sensitivity estimated with a decoding model is a lower bound on the true sensitivity, by showing that a decoding model performs better than chance ($\mathrm{E}[-\log p_\theta(s|r)] < \mathrm{E}[-\log p(s)]$), one shows that the sensitivity of *r* to *s* must be larger than 0. However, comparative claims are more difficult to interpret. Even when the sensitivity of *r* to *s*, estimated with a decoding model, is larger than the sensitivity of *r* to *n*, this does not necessarily mean that *r* is truly more sensitive to *s* than to *n* because one might have underestimated the sensitivity of *r* to *n*.

Two perspectives on decoding models can be distinguished. First, the decoding model can be taken as a model of the statistical relation between variables and as a tool for quantifying sensitivity. As more complex models — with sufficient precautions to avoid overfitting — tend to be more flexible in approximating statistical relations than simpler models, this perspective motivates the use of complex machine learning tools such as support vector machines[104] or more unconstrained deep neural networks[105]. A second perspective interprets the decoding model not just as a model of a statistical relation but as a model of the process by which information about *s* is read from *r*

within the brain. Under this perspective, the complexity of decoding models is often deliberately restricted, for instance, to linear models, because it is assumed that linearly decodable features can be read by a single downstream neuron or layer[8,48]. In this perspective, the decoding model is not simply a tool for estimating sensitivity but also a model of the functionality of r that aims to uncover the neural code that is being used in the brain (Supplementary Box 1).

Encoding models

An encoding model predicts r from s and n (for example, by modelling $p(r|s,n)$). Encoding models can be used to estimate specificity analogously to how decoding models are used to estimate sensitivity. In addition, by determining the role n has in an encoding model of r, invariance to n can be demonstrated (r is invariant to n given that s and n predict no more of r's variability than s alone; Fig. 3c). Examples of encoding models include models of the receptive fields of neurons in the cat visual cortex[31,106], models of shape-tuning in populations of neurons in monkey V4 (ref. 107), and population receptive field models with fMRI of the human visual cortex[32]. General linear models in fMRI analysis[108,109] are also encoding models, as are generalized linear models[110] and linear-non-linear models[111] in the analysis of electrophysiology data.

Analogous to decoding models, the expected logarithmic loss of an encoding model that predicts r from s ($\mathrm{E}[-\log p_\theta(r|s)]$) can be used to estimate the specificity of r to s.

$$\text{Specificity}: \frac{I(s\,;\,r)}{H(r)} = \frac{H(r) - H(r|s)}{H(r)} = \frac{\mathrm{E}[-\log p(r)] - \mathrm{E}[-\log p_\theta(r|s)]}{\mathrm{E}[-\log p(r)]} \quad (9)$$

The lower bound on the expected logarithmic loss of the encoding model corresponds to the upper bound on specificity; when r can perfectly be predicted from s — $\mathrm{E}[-\log p_\theta(r|s)] = 0$ — then specificity is 1. When the expected logarithmic loss is above the baseline predictability of the neural population without knowledge of s ($\mathrm{E}[-\log p(r)]$) — $\mathrm{E}[-\log p_\theta(r|s)] > \mathrm{E}[-\log p(r)]$ — then specificity is larger than 0. The baseline predictability is often assessed by shuffling the relationship between r and s, thereby setting the shuffled r and s to be independent and thus removing any mutual information between r and s in the dataset.

As in the case of decoding models, encoding models can take many forms, from linear models with Gaussian noise to much more complex deep neural network-based models[112], and estimates of specificity established with them are merely lower bounds on the true specificity. Comparisons of specificity estimates, such as the claim that the specificity of r to s is larger than the specificity of r to n, are only as well supported as the assumption that one's encoding model captures all the dependencies between r and s and between r and n.

Encoding model performance can be measured with all the measures that are also available for decoding models. One that is common particularly for encoding models is the proportion of variance explained, which is similar to the information-theoretic quantification of specificity (Fig. 2 and Box 1), except that the variability unexplained by s is measured with the variance of the residual of a model of r given s rather than the entropy:

$$\text{Specificity}: \frac{I(s\,;\,r)}{H(r)} = \frac{H(r) - H(r|s)}{H(r)} = 1 - \frac{H(r|s)}{H(r)} \quad (10)$$

$$\text{Proportion of variance explained}: \quad 1 - \frac{\mathrm{var}(r - m(r\,;s))}{\mathrm{var}(r)}, \quad (11)$$

where $m(r;s)$ is a model that predicts r from s.

Variance is a computationally convenient measure of variability, both for discrete and continuous variables, that can be estimated from small datasets more robustly than entropy but it has limitations. For instance, variance is not sensitive to the shape of probability distributions (such as whether they are skewed or multi-modal).

When s is continuous, an encoding model may also be used to estimate the sensitivity of r to s with the Fisher information ($J(s)$)[113–116], defined as $J(s) = \mathrm{E}[(\frac{d}{ds}\log p(r|s))^2]$. Intuitively, by considering the derivative of a (log-transformed) encoding model ($\frac{d}{ds}\log p(r|s)$), $J(s)$ quantifies locally around a given value of s how much of a difference small changes in s make for r and therefore how informative r is about s. $J(s)$ is closely related to the signal-to-noise ratio (SNR)[117] and to the signal detection theory measure of discriminability (d'), which quantifies how well two neighbouring values of s can be discriminated based on r[116–118]. $J(s)$ and related measures of discriminability (like signal-to-noise ratio and d') are evaluated commonly in analyses of shapes of tuning curves and noise correlations to study properties of the neural code for s[117–124] but also in analyses abstracting away from tuning curves[116,125]. By the Cramér–Rao bound[126,127], which states that the inverse variance of any unbiased decoding model of s from r cannot exceed $J(s)$, $J(s)$ can be used to establish an upper bound on decoding performance (sensitivity) based on an encoding model. Evidence for functionality is provided by demonstrating that measures of how well s can be discriminated given r (based on an encoding model) match measures of how well s is discriminated by b. It was shown, for instance, that the minimal differences in orientation humans could reliably discriminate matched the minimal difference in orientation that could be discriminated based on an encoding model of orientation in V1 (ref. 115).

Representational similarity analysis

Representational similarity analysis (RSA) seeks to build an understanding of low-dimensional geometric properties of high-dimensional neural population activities and their dependence on stimulus features[128]. RSA is related to and builds upon analyses of representational geometry more generally[129–131], including multidimensional scaling[132–134] and principal component analysis[135,136].

In RSA, the geometry of neural representations is captured by computing a representational dissimilarity matrix (RDM). The RDM is constructed by computing dissimilarity in population responses to sets of stimuli. Given a set of N stimuli (S_1, ..., S_N), an $N \times N$ dissimilarity matrix can be constructed, where each entry d_{ij} is calculated by computing the dissimilarity in the neural population response to stimuli S_i and S_j (ref. 128) (Fig. 4b,c). The most used dissimilarity measure is the correlation distance, defined as 1 minus the correlation (ρ) between the population responses to S_i and S_j, computed across neurons or voxels. Alternative dissimilarity measures have been proposed, such as classification accuracy of decoding models, Euclidean distance between values of r[137], or measures based on non-linear transformations of r[138].

Typically, when computing the RDM, each stimulus is presented multiple times and the average neural population response to each stimulus is used to compute dissimilarity. This stimulus-conditioned averaging of the neural population response averages out the non-stimulus-dependent variability. As a result, RDMs tend to poorly reflect the specificity of r to the stimulus set and to more closely reflect its sensitivity.

RDMs can be used to evaluate sensitivity and invariance. If r carries no information about the stimulus, the dissimilarity of neural activity patterns will be the same, independently of the stimulus they were

measured in response to. The resultant RDM would be unstructured and uniformly dominated by noise (Fig. 4b). As such an unstructured RDM suggests zero mutual information between *r* and *s*, it also suggests that sensitivity is 0. Conversely, if a neural population carries information about the stimuli, we expect the RDM to display a structure reflecting the dissimilarity of stimuli. Often, stimuli are drawn from categories with multiple example stimuli present from each category. Dissimilarity of stimuli across category boundaries (such as houses versus faces) indicates sensitivity to the category distinction. Low dissimilarity of stimuli within categories indicates invariance to within-category distinctions (such as variations of the orientation of a face; Fig. 4c). Similarity of RDMs computed from *r* and RDMs computed from *b* can also provide evidence for the functionality of the neural response[139].

Perhaps the most powerful application of RSA lies in comparisons across systems such as the brains of different individuals, different brain areas, brains of different species[140], or even brains and artificial neural networks[141,142]. Yet, such comparisons are not primarily concerned with what evidence establishes that a feature is represented — the focus of this Perspective — but with studying how structures of representations compare across systems.

Tests of statistical dependence

Some analyses do not estimate the amount of information between variables, and therefore do not offer quantitative estimates of the conceptual dimensions of representation. Rather, they test hypotheses about relations of statistical dependence, which suggest that respective dimensions are either satisfied or fail. For instance, initial work demonstrating that the FFA is sensitive to faces showed that fMRI measurements of activity in FFA were significantly higher in conditions in which an image of a face was shown to a participant compared to another object, for example, a house[9]. Of note, this established that FFA carries information about faces without quantifying how much information.

Establishing mutual information between *r* and *s* also establishes that *r* has a larger-than-zero sensitivity and specificity to *s*. That FFA responded to faces irrespective of whether the face is viewed from the front or from the side suggests invariance to viewing angle. That FFA did not respond to other object categories like houses suggests specificity to faces. Yet, such interpretations in terms of invariance and specificity have to be seen with caution as they rest on a null-finding.

Many tests for relations of statistical dependence merely look for a dependence between *s* or *b* and the mean neural response of an isolated element (for example, an individual neuron or a voxel in a fMRI analysis) in a so-called univariate analysis. Such tests cannot pick up on the information carried by activity patterns distributed across neural populations. To pick up on such patterns, a multivariate pattern analysis in terms of correlation or more complex encoding or decoding models is needed[7,29].

Adaptation

The conceptual dimensions of representation may also be evaluated using the phenomenon of adaptation. Upon repeated presentation of an unchanging feature, the response magnitude of neurons[143,144] or brain regions[145,146] that are sensitive to the feature typically decreases. This adaptation is disrupted when the feature changes. Hence, the established computational or physiological mechanism of adaptation can be used to probe the characteristics of a representation. When adaptation is disrupted by changes in *s*, *r* is sensitive to *s*. When adaptation persists across changes in *n*, *r* is invariant to *n*; as far as *r* is concerned, stimuli that differ in *n* are still equivalent.

To illustrate, fMRI activity in the lateral occipital complex adapts to presentations of the same object[145]. Changes in illumination or viewpoint more strongly disrupt this adaptation than changes in position or size of the object. That is, the lateral occipital complex shows some invariance to the position or size of an object, but it is less invariant (more sensitive) to its illumination and viewpoint.

In summary, information theory may be applied in neural data analysis, which affords the most straightforward interpretation in terms of our framework. Yet, we have also explained how other common analysis methods such as linear correlation, decoding and encoding models, tests of statistical dependence, representational similarity analysis, and findings about adaptation may be used to evaluate the conceptual dimensions of representation of sensitivity, specificity, invariance and functionality.

Canonical examples

In our view, neuroscientific research about representations has always been guided implicitly by the framework we have laid out. Our framework is supposed to capture, in a systematic and unified way, what researchers have been doing all along. We look at a few examples of lines of research. These examples were chosen arbitrarily because they have received high attention from researchers and been involved in developing and shaping the tools and analyses with which representations in the brain are being studied. We do not offer exhaustive reviews, but focus on illustrating how tests of the conceptual dimensions of representation combine to form comprehensive bodies of evidence. A single study may not address all conceptual dimensions but, as a line of research matures, evidence across all of them tends to accumulate.

Orientation of visual elements

Research about the visual representation of orientation began with characterizing the response profiles of cells in V1 (refs. 31,106,147). Researchers used encoding models in terms of tuning curves, which estimate the level of activity (and variance) as a function of features, for example, orientation, to evaluate specificity to orientation[148,149]. While the V1 population of neurons is sensitive not only to orientation but also to other features such as contrast, spatial frequency and spatial phase of oriented gratings, some subpopulations, such as complex cells, display invariance at least to spatial phase while maintaining sensitivity to orientation[150]. In both simple and complex cells, the preferred orientation and width (but not the height) of tuning curves is invariant to contrast[151,152] and temporal frequency[153]. Yet, at least the contrast invariance of the width of tuning may depend on adaptation to contrast[154].

On the level of population responses, sensitivity of V1 to orientation has been further characterized with decoding models (both linear and non-linear)[155–158]. Berens et al.[155] showed that, even though individual neurons are not invariant to contrast, at the population level, orientation could be decoded invariant to contrast. Chen et al.[156] evaluated conditional specificity by showing that, even for very low contrast levels, the signal-to-noise ratio for responses to orientation was large.

That V1 response patterns are functional with respect to orientation has been tested by predicting *b* to oriented stimuli from V1 responses. The perceived orientation reported by monkeys with saccadic eye movements can be predicted above chance from single V1 neurons, even in trials where the stimulus contained no oriented signal,

suggesting that the orientation percept is caused by random variations of V1 responses[159]. At the level of V1 population responses, classification of orientations by monkeys can be predicted, using a deep neural network decoder, even when the same stimulus is presented multiple times across trials[160], again suggesting that behavioural responses are caused by random fluctuations in the V1 response across trials.

Numerosity

The numerosity of a collection of things is the number or cardinality of how many things there are in a collection.

On the one hand, initial observations of deficits in the processing of numerosities owing to lesions of the human parietal cortex suggested that the parietal cortex played a causal role in the processing of numerosities. That is, they suggested that there is a parietal representation of numerosity that is functional[161]. Intervention studies have subsequently confirmed this finding of functionality. The repetitive stimulation of parietal areas in humans with transcranial magnetic stimulation impaired the processing of numerosity[162], and pharmacological inactivation of numerosity-sensitive neurons in the monkey's parietal cortex stopped behavioural responses to numerosity[163]. In addition, early fMRI work with humans showed that parietal activity statistically depended on numerical distance in a task of comparing numerosities[164], and neurophysiology with monkeys confirmed that the firing rate of parietal neurons statistically depended on numerosity[165]. The statistical dependence on numerosity implies that parietal neural responses carry information about and are sensitive to numerosity, while a lack of dependence on other factors suggests specificity to numerosity.

On the other hand, invariance has been an ongoing source of scepticism about whether numerosity is really represented in the parietal cortex (rather than a generic quantity that is a mixture of multiple quantitative features). We see here that the conceptual dimensions of representation cannot be used only in support of but also in criticism of claims about representations. Invariance of parietal cortex representations has been demonstrated for features, including low-level visual features[56,165], symbolic and non-symbolic[164], and visual versus auditory presentations of numerosities[166], suggesting that it is numerosity specifically that is being represented. However, controversy remains, with evidence suggesting a lack of invariance, especially to spatial features such as size. In one study[167], encoding models with tuning to preferred numerosities were fit to fMRI data and the preferred numerosity depended on item size. Another study[168] found that neural activity in overlapping areas of the parietal cortex depended on both numerosity and the size of Arabic numerals used to present the numerosity. This lack of invariance suggests that, perhaps, not numerosity specifically but rather a mixed feature that combines spatial size and numerical size, is represented. These conflicting findings could be reconciled if there is a distributed population code for numerosity and related quantities[169] that supports invariant readouts of numerosity by downstream neural processes only in task-specific settings. When numerosity is not task relevant, a mixed quantitative feature may be represented: only when it is task relevant may neurons be tuned to represent numerosity specifically[49].

Spatial location

Place cells are cells in the hippocampus that respond with increases in spiking activity when the animal is in a particular location in the environment, and were discovered in 1971 by O'Keefe and Dostrovsky[170]. The sensitivity of place cells to spatial location is conditional on the presence of visual cues but is invariant to any particular visual cue, such as oriented gratings moving in the rat's visual field[170], as well as further environmental cues, including olfactory and auditory cues[171–173]. As place cells do not fire in response to modulations of such environmental cues, they are also specific to spatial location.

Across the complete population of cells in the hippocampus, the animal's position for all investigated locations can be inferred[174], suggesting that there are not just place cells sensitive to a few specific locations but that the population response is sensitive to location across the animal's environment. Some place cells have specificity to spatial location in a relative reference frame defined by reward location or landmarks that is invariant to absolute location in the experimental environment, whereas others show specificity to absolute location and invariance to changes in reward location and landmarks[175]. Sensitivity to space across the environment in an absolute reference frame has been taken to suggest that the animal represents a map of its environment[176].

Stimulation of place cells that are sensitive to locations previously associated with reward or starting position induced behaviour appropriate for those locations, even when the rat was not located there[177]. Rewarding stimulation of the median forebrain bundle correlated with spontaneous place cell activity during sleep to induce associations between specific locations, and reward also induced goal-directed behaviour by the rat towards those locations when placed in the experimental environment again after awakening[178]. Both interventions suggest that place cells play a causal role in producing location-specific behaviour and are therefore functional as representations of spatial location.

In all three examples — the representation of orientation in V1, the representation of numerosity in the parietal cortex, and the representation of spatial location in the hippocampus — all four conceptual dimensions of representation have been addressed in the literature. Each exemplifies mature lines of research. Still, controversies may remain. Individual neurons in V1 lack invariance to features such as contrast. The population response affords invariant readouts of orientation, but the mechanism by which downstream processes read information about orientation from V1 requires further investigation. The invariance of parietal representations of numerosity to spatial size remains controversial as well. Yet, while there may be disagreement, for instance, about the extent to which invariance is satisfied in some of these cases, there is no disagreement over whether invariance is relevant to a feature being represented. We believe these examples show that the evidence relevant to claims about representation in neuroscience is systematized well by our framework.

Extensions of our framework of conceptual dimensions of representation

For ease of exposition, simplifications have been made in our discussion of the framework. We want to acknowledge some of the most notable limitations and suggest extensions.

First, there is more to say about r. Implicitly, we have treated r as a maximally detailed measure of activity across a population of neurons. However, some work is concerned more specifically with modelling which aspect of a population response is relevant for representing s. A classic question concerns whether the timing of spikes or merely the firing rate matters[14,30,179]. A related question is how a single population of neurons displays mixed selectivity (where different aspects of r may represent different features s_1, s_2, ...)[36,37]. Our framework can also be applied to models of the relevant aspect of r (Supplementary Box 3 explains its application for this purpose).

Second, there is more to say about values of s and how they relate to values of r. One may not only want to assign s as the content of r but also model what value of s is represented by a particular activity pattern of r. That is, one may want to be able to read the neural code (Supplementary Box 1 explains the application of our framework to models of the neural code).

Besides the relaxation of these two simplifications, there are a few more caveats we want to consider.

Neuroscientific work on representations is often an extension of behavioural psychological research. Behavioural work identifies which features are represented (it identifies s), and neuroscientific work localizes these representations in the brain (it identifies r). Where neural data are particularly difficult to collect, for instance, in social cognition or developmental psychology, which focuses on humans, great apes or children, most research on representations is based on behavioural data alone. Without estimates of neural variability, specificity cannot be applied. Yet, models based on behavioural data aim to identify representations that are sensitive to s, they reveal interferences between representations (failures of invariance) and, inherently, they also show that a representation is functional by being used in the production of b.

Further, the information-theoretic quantifications of the conceptual dimensions of representation (Fig. 2) reflect not only dependencies between variables but also the distributions of the variables in isolation ($p(s)$, $p(n)$ and $p(r)$). For instance, naturally distributed stimuli lead to higher estimates of the specificity of spike trains coding for auditory stimuli than unnaturally distributed stimuli[180]. This suggests that the neural code is optimized for the natural statistics of features. However, it is not generally the case that the conceptual dimensions of representation are more informative when evaluated under natural statistics. Often, stimuli are deliberately drawn from unnatural distributions, which allows researchers to tease apart features that tend to covary in natural environments. For example, in the study of representations of numerosity, stimulus distributions are crafted in which features, such as numerosity, item size and item spacing, vary independently as much as possible[181].

Finally, we want to acknowledge that, in this Perspective, the discussion applies our framework only to representations of features of the participant's environment. Elsewhere, we have discussed how our framework applies to representations of uncertainty[182]. Uncertainty is special because of its observer relativity — it is a feature of a belief or representation that an observer has about the world. To cover neuroscientific work in its full breadth, one would also have to consider, for instance, representations of motor commands, goals, values or imaginings.

Concluding remarks

Despite conceptual and terminological ambiguity, we believe there is implicit agreement in neuroscience regarding what is characteristic of representation. Our formal framework aims to make that agreement explicit. It does so by disambiguating and formalizing four types of relation between features, neural responses, and downstream effects of a neural response characteristic of representation that have not been previously teased apart systematically. We propose that researchers use the framework developed in this paper to describe their findings about the representations they investigate — and to explain how their data analysis methods yield estimates of the respective conceptual dimensions of representation. Using common terminology in this way would facilitate communication across research groups, would make it easier to see how different research approaches combine and would afford easier meta-analyses. Few studies provide evidence for all the conceptual dimensions of representation; strong evidence for representation emerges only out of a combination of studies. Following the framework we present in this Perspective makes more salient which evidence is missing from a line of research; it also allows researchers to determine when strong evidence for a representation has emerged in a field.

As measurement techniques develop, datasets get larger and analysis methods get more complex, information-theoretic quantities can be estimated more robustly. Deep neural networks are particularly useful tools that need not be interpreted as models of the brain. In some cases, they may simply be tools for estimating conditional entropies, affording quantitative estimates of the conceptual dimensions of representation for different features and allowing for richer comparisons across studies (see 'Decoding models' and 'Encoding models' for a discussion).

How does our framework relate to philosophical discussions of representation? It is intended to systematize how representations are studied in neuroscience and is not designed to address philosophical questions about what a representation is. However, of course, how representations are studied in neuroscience is highly relevant for philosophical accounts of what a representation is. It will therefore be relevant to consider how philosophical accounts of representation interface with our framework.

Finally, interest has recently arisen in explaining artificial neural networks in representational terms. On the one hand, neuroscience benefits from developments in artificial intelligence research because that research provides new analysis and modelling tools for neuroscience. On the other hand, analysis tools for understanding artificial neural networks are being compared to neuroscientific methodology[183,184]. Our framework is a natural starting point for such a discussion because it encapsulates the methodology that has developed over more than 100 years in neuroscience, and the information-theoretic formalism lends itself naturally to an application to artificial neural networks.

Published online: 20 March 2026

References

1. Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (MIT Press, 1982).
2. Barack, D. L. & Krakauer, J. W. Two views on the cognitive brain. *Nat. Rev. Neurosci.* **22**, 359–371 (2021).
3. Perkel, D. H. & Bullock, T. H. Neural coding. *Neurosci. Res. Program. Bull.* **6**, 221–348 (1968).
4. Baker, B., Lansdell, B. & Kording, K. P. Three aspects of representation in neuroscience. *Trends Cogn. Sci.* **26**, 942–958 (2022).
5. Favela, L. H. & Machery, E. Investigating the concept of representation in the neural and psychological sciences. *Front. Psychol.* **14**, 1165622 (2023).
6. Salzman, C. D., Britten, K. H. & Newsome, W. T. Cortical microstimulation influences perceptual judgements of motion direction. *Nature* **346**, 174–177 (1990).
7. Haxby, J. V. et al. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
8. Kriegeskorte, N. & Diedrichsen, J. Peeling the onion of brain representations. *Annu. Rev. Neurosci.* **42**, 407–432 (2019).
9. Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
10. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
11. Dretske, F. *Knowledge and the Flow of Information* (The MIT Press, 1981).
12. Usher, M. A statistical referential theory of content: using information theory to account for misrepresentation. *Mind Lang.* **16**, 311–334 (2001).
13. Quiroga, R. Q. & Panzeri, S. Extracting information from neuronal populations: information theory and decoding approaches. *Nat. Rev. Neurosci.* **10**, 173–185 (2009).
14. Bialek, W., Rieke, F., de Ruyter van Steveninck, R. & Warland, D. Reading a neural code. *Science* **252**, 1854–1857 (1991).

15. Bialek, W., DeWeese, M., Rieke, F. & Warland, D. Bits and brains: information flow in the nervous system. *Phys. Stat. Mech. Appl.* **200**, 581–593 (1993).
16. Laughlin, S. Energy as a constraint on the coding and processing of sensory information. *Curr. Opin. Neurobiol.* **11**, 475–480 (2001).
17. Rieke, F., Warland, D. & Bialek, W. Coding efficiency and information rates in sensory neurons. *Europhys. Lett. EPL* **22**, 151–156 (1993).
18. Ince, R. A. A. et al. Tracing the flow of perceptual features in an algorithmic brain network. *Sci. Rep.* **5**, 17681 (2015).
19. Panzeri, S., Harvey, C. D., Piasini, E., Latham, P. E. & Fellin, T. Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. *Neuron* **93**, 491–507 (2017).
20. Pica, G. et al. in *Advances in Neural Information Processing Systems* (eds Guyon, I. et al.) vol. 30 (NIPS Foundation, 2017).
21. Varley, T. F., Sporns, O., Schaffelhofer, S., Scherberger, H. & Dann, B. Information-processing dynamics in neural networks of macaque cerebral cortex reflect cognitive state and behavior. *Proc. Natl. Acad. Sci. USA* **120**, e2207677120 (2023).
22. Borst, A. & Theunissen, F. E. Information theory and neural coding. *Nat. Neurosci.* **2**, 947–957 (1999).
23. Ince, R. A. A. et al. A statistical framework for neuroimaging data analysis based on mutual information estimated via a Gaussian copula. *Hum. Brain Mapp.* **38**, 1541–1573 (2017).
24. Ostwald, D. & Bagshaw, A. P. Information theoretic approaches to functional neuroimaging. *Magn. Reson. Imaging* **29**, 1417–1428 (2011).
25. Panzeri, S., Magri, C. & Logothetis, N. K. On the use of information theory for the analysis of the relationship between neural and imaging signals. *Magn. Reson. Imaging* **26**, 1015–1025 (2008).
26. deCharms, R. C. & Zador, A. Neural representation and the cortical code. *Annu. Rev. Neurosci.* **23**, 613–647 (2000).
27. Zhang, K., Ginzburg, I., McNaughton, B. L. & Sejnowski, T. J. Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. *J. Neurophysiol.* **79**, 1017–1044 (1998).
28. Kragel, P. A., Koban, L., Barrett, L. F. & Wager, T. D. Representation, pattern information, and brain signatures: from neurons to neuroimaging. *Neuron* **99**, 257–273 (2018).
29. Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* **37**, 435–456 (2014).
30. Rieke, F., Warland, D., de Ruyter van Steveninck, R. & Bialek, W. S. *Spikes: Exploring the Neural Code* (The MIT Press, 1999).
31. Hubel, D. H. & Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **148**, 574–591 (1959).
32. Dumoulin, S. O. & Wandell, B. A. Population receptive field estimates in human visual cortex. *NeuroImage* **39**, 647–660 (2008).
33. Chang, L. & Tsao, D. Y. The code for facial identity in the primate brain. *Cell* **169**, 1013–1028.e14 (2017).
34. Dubner, R. & Zeki, S. M. Response properties and receptive fields of cells in an anatomically defined region of the superior temporal sulcus in the monkey. *Brain Res.* **35**, 528–532 (1971).
35. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, 2007).
36. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).
37. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
38. Haueis, P. & Colaço, D. J. Metabolic considerations for cognitive modeling. *Behav. Brain Sci.* https://doi.org/10.1017/S0140525X25103956 (2025).
39. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410 (2011).
40. Poldrack, R. A. Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* **10**, 59–63 (2006).
41. Poldrack, R. A. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron* **72**, 692–697 (2011).
42. Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* **7**, 358–366 (2006).
43. Faisal, A. A., Selen, L. P. J. & Wolpert, D. M. Noise in the nervous system. *Nat. Rev. Neurosci.* **9**, 292–303 (2008).
44. Logothetis, N. K., Pauls, J. & Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* **5**, 552–563 (1995).
45. 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).
46. Zeki, S. The representation of colours in the cerebral cortex. *Nature* **284**, 412–418 (1980).
47. Chawla, D., Rees, G. & Friston, K. J. The physiological basis of attentional modulation in extrastriate visual areas. *Nat. Neurosci.* **2**, 671–676 (1999).
48. DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007).
49. Castaldi, E., Piazza, M., Dehaene, S., Vignaud, A. & Eger, E. Attentional amplification of neural codes for number independent of other quantities along the dorsal visual stream. *eLife* **8**, e45160 (2019).
50. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (John Wiley & Sons, 2006).
51. Burge, T. Origins of perception. *Disputatio* **4**, 1–38 (2010).
52. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
53. Friston, K. J. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).
54. Helmholtz, H. *Handbuch Der Physiologischen Optik* (Leopold Voss, 1867).
55. Hohwy, J. *The Predictive Mind* (Oxford University Press, 2013).
56. Nieder, A., Freedman, D. & Miller, E. K. Representation of the quantity of visual items in the primate prefrontal cortex. *Science* **297**, 1708–1711 (2002).
57. Ritchie, J. B., Kaplan, D. M. & Klein, C. Decoding the brain: neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *Br. J. Philos. Sci.* **70**, 581–607 (2019).
58. Shadlen, M. N. & Newsome, W. T. Noise, neural codes and cortical organization. *Curr. Opin. Neurobiol.* **4**, 569–579 (1994).
59. Brette, R. Is coding a relevant metaphor for the brain? *Behav. Brain Sci.* **42**, e215 (2019).
60. Jones, I. S. & Kording, K. P. Quantifying the role of neurons for behavior is a mediation question. *Behav. Brain Sci.* **42**, e233 (2019).
61. Pearl, J. *Causality: Models, Reasoning, and Inference* (Cambridge University Press, 2009).
62. Woodward, J. *Making Things Happen: A Theory of Causal Explanation* (Oxford University Press, 2003).
63. Bressler, S. L. & Seth, A. K. Wiener–Granger causality: a well established methodology. *NeuroImage* **58**, 323–329 (2011).
64. Friston, K. J., Harrison, L. & Penny, W. Dynamic causal modelling. *NeuroImage* **19**, 1273–1302 (2003).
65. Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424 (1969).
66. Valdes-Sosa, P. A., Roebroeck, A., Daunizeau, J. & Friston, K. Effective connectivity: Influence, causality and biophysical modeling. *NeuroImage* **58**, 339–361 (2011).
67. Weichwald, S. & Peters, J. Causality in cognitive neuroscience: concepts, challenges, and distributional robustness. *J. Cogn. Neurosci.* **33**, 226–247 (2021).
68. Bastos, A. M. & Schoffelen, J.-M. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Front. Syst. Neurosci.* **9**, 175 (2016).
69. Friston, K. J. Functional and effective connectivity in neuroimaging: a synthesis. *Hum. Brain Mapp.* **2**, 56–78 (1994).
70. Luppi, A. I. et al. A synergistic core for human brain evolution and cognition. *Nat. Neurosci.* **25**, 771–782 (2022).
71. Stephan, K. E. & Friston, K. J. in *Encyclopedia of Neuroscience* (ed Squire, L. R.) 391–397 (Elsevier, 2009).
72. Celotto, M. et al. in *Advances in Neural Information Processing Systems* (eds Oh, A. et al.) vol. 36 (NeurIPS, 2023).
73. Lemke, S. M., Celotto, M., Maffulli, R., Ganguly, K. & Panzeri, S. Information flow between motor cortex and striatum reverses during skill learning. *Curr. Biol.* **34**, 1831–1843.e7 (2024).
74. Purushothaman, G. & Bradley, D. C. Neural population code for fine perceptual decisions in area MT. *Nat. Neurosci.* **8**, 99–106 (2005).
75. Rossi-Pool, R., Zainos, A., Alvarez, M., Diaz-deLeon, G. & Romo, R. A continuum of invariant sensory and behavioral-context perceptual coding in secondary somatosensory cortex. *Nat. Commun.* **12**, 2000 (2021).
76. Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S. & Movshon, J. A. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis. Neurosci.* **13**, 87–100 (1996).
77. Wilming, N., Murphy, P. R., Meyniel, F. & Donner, T. H. Large-scale dynamics of perceptual decision information across human cortex. *Nat. Commun.* **11**, 5109 (2020).
78. Bradley, D. C., Chang, G. C. & Andersen, R. A. Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature* **392**, 714–717 (1998).
79. Hanks, T. D., Ditterich, J. & Shadlen, M. N. Microstimulation of macaque area LIP affects decision-making in a motion discrimination task. *Nat. Neurosci.* **9**, 682–689 (2006).
80. Newsome, W. & Pare, E. A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *J. Neurosci.* **8**, 2201–2211 (1988).
81. Beckers, G. & Homberg, V. Cerebral visual motion blindness: transitory akinetopsia induced by transcranial magnetic stimulation of human area V5. *Proc. R. Soc. Lond. B Biol. Sci.* **249**, 173–178 (1992).
82. Moore, T. & Armstrong, K. M. Selective gating of visual signals by microstimulation of frontal cortex. *Nature* **421**, 370–373 (2003).
83. Nemenman, I., Bialek, W. & de Ruyter van Steveninck, R. Entropy and information in neural spike trains: progress on the sampling problem. *Phys. Rev. E* **69**, 056111 (2004).
84. Victor, J. D. Approaches to information-theoretic analysis of neural activity. *Biol. Theory* **1**, 302–316 (2006).
85. Saproo, S. & Serences, J. T. Spatial attention improves the quality of population codes in human visual cortex. *J. Neurophysiol.* **104**, 885–895 (2010).
86. Serences, J., Saproo, S., Scolari, M., Ho, T. & Muftuler, L. Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. *NeuroImage* **44**, 223–231 (2009).
87. Guggenmos, M. et al. Spatial attention enhances object coding in local and distributed representations of the lateral occipital complex. *NeuroImage* **116**, 149–157 (2015).
88. Schyns, P. G., Thut, G. & Gross, J. Cracking the code of oscillatory activity. *PLoS Biol.* **9**, e1001064 (2011).
89. Caballero-Gaudes, C. et al. Mapping interictal epileptic discharges using mutual information between concurrent EEG and fMRI. *NeuroImage* **68**, 248–262 (2013).
90. Gross, J. et al. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biol.* **11**, e1001752 (2013).
91. Alemi, A. A., Fischer, I., Dillon, J. V. & Murphy, K. Deep variational information bottleneck. In *International Conference on Learning Representations* (2017).

92. Álvarez Chaves, M., Gupta, H. V., Ehret, U. & Guthke, A. On the accurate estimation of information-theoretic quantities from multi-dimensional sample data. *Entropy* **26**, 387 (2024).
93. Kraskov, A., Stögbauer, H. & Grassberger, P. Estimating mutual information. *Phys. Rev. E* **69**, 066138 (2004).
94. Piga, A., Font-Pomarol, L., Sales-Pardo, M. & Guimerà, R. Bayesian estimation of information-theoretic metrics for sparsely sampled distributions. *Chaos Solitons Fractals* **180**, 114564 (2024).
95. Nagel, D., Diez, G. & Stock, G. Accurate estimation of the normalized mutual information of multidimensional data. *J. Chem. Phys.* **161**, 054108 (2024).
96. Tafazoli, S. et al. Emergence of transformation-tolerant representations of visual objects in rat lateral extrastriate cortex. *eLife* **6**, e22794 (2017).
97. Reshef, D. N. et al. Detecting novel associations in large data sets. *Science* **334**, 1518–1524 (2011).
98. Abbott, L. F., Rolls, E. T. & Tovee, M. J. Representational capacity of face coding in monkeys. *Cereb. Cortex* **6**, 498–505 (1996).
99. Hung, C. P., Kreiman, G., Poggio, T. & DiCarlo, J. J. Fast readout of object identity from macaque inferior temporal cortex. *Science* **310**, 863–866 (2005).
100. Hassabis, D. et al. Decoding neuronal ensembles in the human hippocampus. *Curr. Biol.* **19**, 546–554 (2009).
101. Miller, L. M. & Recanzone, G. H. Populations of auditory cortical neurons can accurately encode acoustic space across stimulus intensity. *Proc. Natl. Acad. Sci. USA* **106**, 5931–5935 (2009).
102. Cox, D. D. & Savoy, R. L. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* **19**, 261–270 (2003).
103. Friston, K. J. et al. Bayesian decoding of brain images. *NeuroImage* **39**, 181–205 (2008).
104. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
105. Livezey, J. A. & Glaser, J. I. Deep learning approaches for neural decoding across architectures and recording modalities. *Brief. Bioinform.* **22**, 1577–1591 (2021).
106. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154 (1962).
107. Pasupathy, A. & Connor, C. E. Population coding of shape in area V4. *Nat. Neurosci.* **5**, 1332–1338 (2002).
108. Friston, K. J., Jezzard, P. & Turner, R. Analysis of functional MRI time-series. *Hum. Brain Mapp.* **1**, 153–171 (1994).
109. Friston, K. J. et al. Analysis of fMRI time-series revisited. *NeuroImage* **2**, 45–53 (1995).
110. Paninski, L., Pillow, J. & Lewi, J. Statistical models for neural encoding, decoding, and optimal stimulus design. *Prog. Brain Res.* **165**, 493–507 (2007).
111. Aljadeff, J., Lansdell, B. J., Fairhall, A. L. & Kleinfeld, D. Analysis of neuronal spike trains, deconstructed. *Neuron* **91**, 221–259 (2016).
112. Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).
113. Brunel, N. & Nadal, J.-P. Mutual information, Fisher information, and population coding. *Neural Comput.* **10**, 1731–1757 (1998).
114. Fisher, R. A. Theory of statistical estimation. *Math. Proc. Camb. Philos. Soc.* **22**, 700–725 (1925).
115. Paradiso, M. A. A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biol. Cybern.* **58**, 35–49 (1988).
116. Wei, X.-X. & Stocker, A. A. Mutual information, Fisher information, and efficient coding. *Neural Comput.* **28**, 305–326 (2016).
117. Seung, H. S. & Sompolinsky, H. Simple models for reading neuronal population codes. *Proc. Natl. Acad. Sci. USA* **90**, 10749–10753 (1993).
118. Averbeck, B. B. & Lee, D. Effects of noise correlations on information encoding and decoding. *J. Neurophysiol.* **95**, 3633–3644 (2006).
119. Bejjanki, V. R., Beck, J. M., Lu, Z.-L. & Pouget, A. Perceptual learning as improved probabilistic inference in early sensory areas. *Nat. Neurosci.* **14**, 642–648 (2011).
120. Haefner, R. M. & Bethge, M. in *Advances in Neural Information Processing Systems* (eds Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R. & Culotta, A.) vol. 23 (NeurIPS, 2010).
121. Pouget, A., Deneve, S., Ducom, J.-C. & Latham, P. E. Narrow versus wide tuning curves: what's best for a population code? *Neural Comput.* **11**, 85–90 (1999).
122. Rumyantsev, O. I. et al. Fundamental bounds on the fidelity of sensory cortical coding. *Nature* **580**, 100–105 (2020).
123. Yoon, H. & Sompolinsky, H. in *Advances in Neural Information Processing Systems* (eds Kearns, M., Solla, S. & Cohn, D.) vol. 11 (NeurIPS, 1998).
124. Zhou, J., Duong, L. R. & Simoncelli, E. P. A unified framework for perceived magnitude and discriminability of sensory stimuli. *Proc. Natl. Acad. Sci. USA* **121**, e2312293121 (2024).
125. Kriegeskorte, N. & Wei, X.-X. Neural tuning and representational geometry. *Nat. Rev. Neurosci.* **22**, 703–718 (2021).
126. Cramér, H. *Mathematical Methods of Statistics* (Princeton University Press, 1946).
127. Rao, C. R. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **37**, 81–91 (1945).
128. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis — connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
129. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).
130. Shepard, R. N. Multidimensional scaling, tree-fitting, and clustering. *Science* **210**, 390–398 (1980).
131. Umakantha, A. et al. Bridging neuronal correlations and dimensionality reduction. *Neuron* **109**, 2740–2754.e12 (2021).
132. Hasselmo, M. E., Rolls, E. T. & Baylis, G. C. The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behav. Brain Res.* **32**, 203–218 (1989).
133. Young, M. P. et al. Non-metric multidimensional scaling in the analysis of neuroanatomical connection data and the organization of the primate cortical visual system. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **348**, 281–308 (1995).
134. Young, M. P. & Yamane, S. Sparse population coding of faces in the inferotemporal cortex. *Science* **256**, 1327–1331 (1992).
135. 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).
136. Nicolelis, M. A. L., Baccala, L. A., Lin, R. C. S. & Chapin, J. K. Sensorimotor encoding by synchronous neural ensemble activity at multiple levels of the somatosensory system. *Science* **268**, 1353–1358 (1995).
137. Walther, A. et al. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage* **137**, 188–200 (2016).
138. Lin, B. & Kriegeskorte, N. The topology and geometry of neural representations. *Proc. Natl. Acad. Sci. USA* **121**, e2317881121 (2024).
139. Op De Beeck, H., Wagemans, J. & Vogels, R. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat. Neurosci.* **4**, 1244–1252 (2001).
140. Kriegeskorte, N. et al. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–1141 (2008).
141. Li, Z. et al. in *Advances in neural information processing systems* (eds Wallach, H. et al.) vol. 32 (NeurIPS, 2019).
142. Yamins, D. L., Hong, H., Cadieu, C. & DiCarlo, J. J. in *Advances in neural information processing systems* (eds Burges, C. J., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K. Q.) vol. 26 (NeurIPS, 2013).
143. Baylis, G. C. & Rolls, E. T. Responses of neurons in the inferior temporal cortex in short term and serial recognition memory tasks. *Exp. Brain Res.* **65**, 614–622 (1987).
144. Sobotka, S. & Ringo, J. L. Stimulus specific adaptation in excited but not in inhibited cells in inferotemporal cortex of Macaque. *Brain Res.* **646**, 95–99 (1994).
145. Grill-Spector, K. et al. Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron* **24**, 187–203 (1999).
146. Weigelt, S., Muckli, L. & Kohler, A. Functional magnetic resonance adaptation in visual neuroscience. *Rev. Neurosci.* **19**, 363–380 (2008).
147. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195**, 215–243 (1968).
148. Sompolinsky, H. & Shapley, R. New perspectives on the mechanisms for orientation selectivity. *Curr. Opin. Neurobiol.* **7**, 514–522 (1997).
149. Victor, J. D., Purpura, K., Katz, E. & Mao, B. Population encoding of spatial frequency, orientation, and color in macaque V1. *J. Neurophysiol.* **72**, 2151–2166 (1994).
150. De Valois, R. L., Albrecht, D. G. & Thorell, L. G. Spatial frequency selectivity of cells in macaque visual cortex. *Vis. Res.* **22**, 545–559 (1982).
151. Anderson, J. S., Lampl, I., Gillespie, D. C. & Ferster, D. The contribution of noise to contrast invariance of orientation tuning in cat visual cortex. *Science* **290**, 1968–1972 (2000).
152. Hansel, D. & van Vreeswijk, C. How noise contributes to contrast invariance of orientation tuning in cat visual cortex. *J. Neurosci.* **22**, 5118–5128 (2002).
153. Moore, B. D., Alitto, H. J. & Usrey, W. M. Orientation tuning, but not direction selectivity, is invariant to temporal frequency in primary visual cortex. *J. Neurophysiol.* **94**, 1336–1345 (2005).
154. Nowak, L. G. & Barone, P. Contrast adaptation contributes to contrast-invariance of orientation tuning of primate V1 cells. *PLoS One* **4**, e4781 (2009).
155. Berens, P. et al. A fast and simple population code for orientation in primate V1. *J. Neurosci.* **32**, 10618–10626 (2012).
156. Chen, Y., Geisler, W. S. & Seidemann, E. Optimal decoding of correlated neural population responses in the primate visual cortex. *Nat. Neurosci.* **9**, 1412–1420 (2006).
157. Chen, Y., Geisler, W. S. & Seidemann, E. Optimal temporal decoding of neural population responses in a reaction-time visual detection task. *J. Neurophysiol.* **99**, 1366–1379 (2008).
158. 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**, 239–245 (2011).
159. Nienborg, H. & Cumming, B. G. Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex. *J. Neurosci.* **34**, 3579–3585 (2014).
160. 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).
161. Dehaene, S. & Cohen, L. Towards an anatomical and functional model of number processing. *Math. Cogn.* **1**, 83–120 (1995).
162. Dormal, V., Andres, M. & Pesenti, M. Dissociation of numerosity and duration processing in the left intraparietal sulcus: a transcranial magnetic stimulation study. *Cortex* **44**, 462–469 (2008).
163. Sawamura, H., Shima, K. & Tanji, J. Deficits in action selection based on numerical information after inactivation of the posterior parietal cortex in monkeys. *J. Neurophysiol.* **104**, 902–910 (2010).
164. Pinel, P. et al. Event-related fMRI analysis of the cerebral circuit for number comparison. *NeuroReport* **10**, 1473–1479 (1999).
165. Nieder, A. & Miller, E. K. A parieto-frontal network for visual numerical information in the monkey. *Proc. Natl. Acad. Sci. USA* **101**, 7457–7462 (2004).
166. Nieder, A. Supramodal numerosity selectivity of neurons in primate prefrontal and posterior parietal cortices. *Proc. Natl. Acad. Sci. USA* **109**, 11860–11865 (2012).

167. Harvey, B. M., Klein, B. P., Petridou, N. & Dumoulin, S. O. Topographic representation of numerosity in the human parietal cortex. *Science* **341**, 1123–1126 (2013).
168. Pinel, P., Piazza, M., Le Bihan, D. & Dehaene, S. Distributed and overlapping cerebral representations of number, size, and luminance during comparative judgments. *Neuron* **41**, 983–993 (2004).
169. Tudusciuc, O. & Nieder, A. Neuronal population coding of continuous and discrete quantity in the primate posterior parietal cortex. *Proc. Natl. Acad. Sci. USA* **104**, 14513–14518 (2007).
170. O'Keefe, J. & Dostrovsky, J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**, 171–175 (1971).
171. Battaglia, F. P., Sutherland, G. R. & McNaughton, B. L. Local sensory cues and place cell directionality: additional evidence of prospective coding in the hippocampus. *J. Neurosci.* **24**, 4541–4550 (2004).
172. Jeffery, K. J. & O'Keefe, J. M. Learned interaction of visual and idiothetic cues in the control of place field orientation. *Exp. Brain Res.* **127**, 151–161 (1999).
173. Sharp, P., Kubie, J. & Muller, R. Firing properties of hippocampal neurons in a visually symmetrical environment: contributions of multiple sensory cues and mnemonic processes. *J. Neurosci.* **10**, 3093–3105 (1990).
174. Wilson, M. A. & McNaughton, B. L. Dynamics of the hippocampal ensemble code for space. *Science* **261**, 1055–1058 (1993).
175. Gothard, K. M., Skaggs, W. E., Moore, K. M. & McNaughton, B. L. Binding of hippocampal CA1 neural activity to multiple reference frames in a landmark-based navigation task. *J. Neurosci.* **16**, 823–835 (1996).
176. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Clarendon Press, 1978).
177. Robinson, N. T. M. et al. Targeted activation of hippocampal place cells drives memory-guided spatial behavior. *Cell* **183**, 1586–1599.e10 (2020).
178. De Lavilléon, G., Lacroix, M. M., Rondi-Reig, L. & Benchenane, K. Explicit memory creation during sleep demonstrates a causal role of place cells in navigation. *Nat. Neurosci.* **18**, 493–495 (2015).
179. London, M., Roth, A., Beeren, L., Häusser, M. & Latham, P. E. Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. *Nature* **466**, 123–127 (2010).
180. Rieke, F., Bodnar, D. A. & Bialek, W. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. R. Soc. Lond. B Biol. Sci.* **262**, 259–265 (1995).
181. 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).
182. Walker, E. Y. et al. Studying the neural representations of uncertainty. *Nat. Neurosci.* **26**, 1857–1867 (2023).
183. Ivanova, A. A., Hewitt, J. & Zaslavsky, N. Probing artificial neural networks: insights from neuroscience. *Prepreint at* http://arxiv.org/abs/2104.08197 (2021).
184. Lindsay, G. W. & Bau, D. Testing methods of neural systems understanding. *Cogn. Syst. Res.* **82**, 101156 (2023).

Acknowledgements
F.M. is supported by an ERC grant 948105-NEURAL-PROB.

Author contributions
All authors contributed substantially to the development of our framework and to discussion of the content. S.P. wrote the paper with contributions from E.Y.W., D.L.B. and J.L. S.P., R.N.D., F.M. and W.J.M. revised and edited the manuscript.

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/s41583-026-01030-8.

Peer review information *Nature Reviews Neuroscience* thanks the anonymous reviewers for their contribution to the peer review of this work.

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