

# Making models disagree to learn how brains compute

Tal Golan<sup>1,2,3</sup>✉, Heiko H. Schütt<sup>4</sup> & Nikolaus Kriegeskorte<sup>5,6,7,8</sup>✉

## Abstract

Computational hypotheses about brain information processing can be expressed in neural network models. Neuroscientists have begun to compare such models in terms of their alignment with neural and behavioural data. The high parametric capacity of these models is essential to their ability to capture cognitive processes but also enables them to approximate arbitrary functions, making distinct models difficult to discriminate experimentally. Model comparisons using stimuli sampled from the training distribution often fail to reveal differences. This challenge can be met by optimizing stimulus sets for model discrimination and by leveraging out-of-distribution generalization as a severe test. This Review explains the emerging methods for optimizing stimuli to adjudicate among neural network models. These methods seek stimulus sets that are controversial among the models in that they make the models disagree in their predictions of the experimental data. We discuss the choices researchers must make, including a prior over candidate stimuli (such as naturalistic images), a measure of the power to discriminate among alternative models and a procedure for selecting or synthesizing stimulus sets that maximize model-comparison power. Historically, researchers have chosen either natural or artificial stimuli for a given study, prioritizing ecological validity or model-comparison power, respectively. Tempered by a prior, controversial stimuli offer a synthesis of these classical approaches, combining the greater ecological validity of naturalistic stimuli with the greater power for model comparison enabled by artificial stimuli. We offer a unified perspective on current work, drawing connections to Bayesian optimal experimental design.

## Sections

Introduction

Challenges in comparing parameter-rich models

Generating stimulus sets that make models disagree

Controversial stimuli in practice

Outlook

<sup>1</sup>Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Be'er-Sheva, Israel.

<sup>2</sup>The School of Brain Sciences and Cognition, Ben-Gurion University of the Negev, Be'er-Sheva, Israel. <sup>3</sup>Data Science Research Center, Ben-Gurion University of the Negev, Be'er-Sheva, Israel. <sup>4</sup>Department of Behavioural and Cognitive Sciences, Université du Luxembourg, Esch-sur-Alzette, Luxembourg. <sup>5</sup>Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA. <sup>6</sup>Department of Psychology, Columbia University, New York, NY, USA.

<sup>7</sup>Department of Neuroscience, Columbia University, New York, NY, USA. <sup>8</sup>Department of Electrical Engineering, Columbia University, New York, NY, USA. ✉e-mail: [golan.neuro@bgu.ac.il](mailto:golan.neuro@bgu.ac.il); [n.kriegeskorte@columbia.edu](mailto:n.kriegeskorte@columbia.edu)

## Introduction

To learn how brains compute, neuroscience must compare alternative theories of brain computation in light of experimental data. Computational theories can be fully specified and rendered empirically testable by implementation in models that simulate the information-processing function investigated<sup>1–10</sup>. One approach to the experimental evaluation of such models is to present samples of natural stimuli to animals or human participants in neurophysiological and behavioural experiments<sup>11</sup>. The same stimuli can be presented to models so as to compare their internal representations and behaviour with the neural representations and behaviour of the participants. Using natural stimuli to evaluate models ensures that the model evaluation is ecologically valid: ultimately, models should explain neural activity and behaviour under natural conditions, so evaluation on natural stimuli is essential<sup>4,12–14</sup>. However, if a model perfectly captures the neural mechanism that produces the neural and behavioural responses, it ought to make good predictions for all stimuli. Researchers therefore should test models also on non-natural stimuli. Including non-natural stimuli also gives researchers greater flexibility to choose stimuli that afford good power for model comparison<sup>15</sup>.

Historically, psychophysicists and neuroscientists seeking to evaluate computational models have often relied on hand-crafted artificial stimuli, designed to provide critical tests of model predictions. Such an approach is feasible when the models are simple enough for researchers to design the stimuli. However, complex perceptual, cognitive and motor control functions require much knowledge about the world, which a model that can perform the task must have the capacity to store. For example, visual recognition of a wide range of objects requires knowledge of the appearance of these objects. The uncomfortable consequence of this need for knowledge is that models must combine hypothetical cognitive and neuroscientific mechanisms with high model capacity and machine learning. Neural network models can express computational hypotheses while also providing the required capacity to store the knowledge needed to capture intelligent behaviour<sup>5,16</sup>. However, the complexity of neural network models makes it difficult to hand-craft stimulus sets that enable powerful model comparisons. The field's reliance on natural stimuli can be motivated in terms of ecological validity but, given the difficulty of designing critical experiments for neural network models, it is also a choice of convenience.

Rather than choosing either artificial or natural stimuli, neuroscience can now leverage optimization techniques to design experimental stimuli for more powerful comparisons among neural network models. Stimuli optimized for model comparison will be driven by model predictions and will not, in general, conform to the statistics of natural stimuli. However, when the models are neural networks trained on natural images, stimulus sets optimized for model comparison reflect the knowledge the models have acquired about the natural world and therefore often take on certain features of natural stimuli. In addition, the stimulus space can be restricted to ensure that model comparisons rely only on natural stimuli or on stimuli considered sufficiently naturalistic to make the experiment ecologically valid. In this way, optimization of stimulus sets for model comparison offers a synthesis of the contrasting established approaches of using artificial or natural stimuli.

In this Review, we discuss how to optimize stimuli to experimentally discriminate high-parametric neural network models, grounded in the normative foundations of experimental design. The stimulus set optimization techniques we describe are relevant for a broad

range of analyses, including probabilistic modelling of behaviour, representational similarity analysis (RSA) and linear encoding models. We begin by discussing the challenges of comparing parameter-rich models. We then examine the key choices that an experimenter must make when optimizing a stimulus set to discriminate models. Next, we review neuroscientific studies that have optimized stimulus sets to elicit disagreement and adjudicate among neural network models. We conclude by discussing outstanding questions and future directions.

## Challenges in comparing parameter-rich models

Recent research on the relationship between artificial neural networks and cortical representations has repeatedly revealed two observations. First, internal activations of parameter-rich neural network models are highly predictive of animal and human neural responses to natural stimuli, even when these models are not directly fitted to neural data<sup>2,3</sup>. Second, multiple distinct neural network models achieve similar levels of predictivity<sup>17,18</sup>. Whereas the former observation suggests that neural network models hold promise for understanding brain computation, the latter casts doubt on the prospects for theoretical progress through iterative model building and model comparison.

These two observations are evident across domains when models are trained and tested on randomly sampled natural stimuli. The 'Brain-Score' project<sup>19,20</sup> benchmarks many neural network models by evaluating their predictions of diverse neural and behavioural datasets using fitted linear mappings from model representations to neural and behavioural responses. Qualitatively distinct models often achieve similar scores on Brain-Score. For example, a range of artificial neural network architectures align equally well with neural recordings from V2, V4 and the inferior temporal cortex<sup>21</sup>. Models designed to implement specific neuroscientific motifs, such as recurrence<sup>22</sup> or V1-like early processing<sup>23</sup>, do not systematically outperform neural networks adopted from machine learning without modification<sup>20</sup> (but see ref. 24 for evidence that recurrent models better explain human ventral-stream representational dynamics). Comparing human functional MRI (fMRI) and hidden-layer artificial neural network activations indicates similar brain alignment of various convolutional neural network architectures trained on object recognition, once linear reweighting of representations is allowed<sup>17</sup>. A recent, larger-scale survey generalizes the finding of equivalent neural prediction accuracy from convolutional to transformer neural network models and to models trained on different visual tasks<sup>18</sup>. In the language domain, also covered by Brain-Score using human neuroimaging data, some models predict the available imaging benchmarks as well as the interparticipant reliability permits<sup>19,20</sup>. Beyond Brain-Score, studies on the alignment of artificial neural networks with human fMRI stimulus representations often report model performance well above chance, but statistical comparisons between models are typically absent, insignificant or uncorrected for multiple testing. Within the domain of neural representations of human faces, studies using primate single-unit recordings<sup>25</sup>, human intracranial recordings<sup>26</sup>, fMRI<sup>27</sup> and human similarity judgements<sup>28</sup> all report tied predictivity scores across two or more distinct computational models. Importantly, even when the models under comparison are not neural network representations of natural stimuli but instead hand-crafted machine-vision image descriptors<sup>29</sup> or human-labelling-based stimulus features<sup>30,31</sup>, evaluations on randomly sampled natural stimuli frequently yield statistically indistinguishable performance levels. This lack of differentiation indicates that the models make predictions that are too similar in their

fit to the data, relative to the sensitivity of the field's experimental and data-analytical methods for model comparison. It remains a matter of debate to what extent the similarity of the models or the sensitivity of the methods is to blame.

Some recent studies argue that indistinguishable models should be considered implementations of a single universal representation. The argument for universality is based on the observation that the most predictive representational features are shared between models<sup>32–34</sup>. Even before this recent appeal to the universal representation hypothesis, there was already a concern that high-parametric deep neural network models may act as mere function approximators that do not provide predictive value beyond a data summary<sup>17,18</sup>. Whether alternative models make empirically distinguishable predictions can be assessed using model-recovery simulations: for each model in turn, researchers imagine that it is the true data-generating mechanism. They then simulate a noisy dataset on the basis of the model's predictions, assuming a realistic noise level. Now the researchers pretend that they have acquired this dataset and do not know which model generated it. They apply a model-comparative analysis to identify the data-generating model. Because the data are simulated, the researchers know the true model and can evaluate their model-comparative inference objectively – by counting how often it correctly recovers the data-generating model. If the model-recovery accuracy is high, then the model predictions are sufficiently distinct to enable adjudication. Using only naturalistic stimuli in the context of vision, model-recovery simulations give a mixed picture. The 'glass half full' interpretation is that they often enable discrimination of models. The 'glass half empty' interpretation is that they also regularly fail to discriminate even models that have strikingly different architectures<sup>21,35</sup>.

We can reconcile the universalist and particularist perspectives by noting that diverse models and brains all need to learn to represent major variance-explaining latent factors of the sensory data. All visual systems, after all, must represent the same visual world. Even a vision and a language model<sup>32</sup> will come to reflect some of the same latent factors that organize the world. For example, consider that forks and knives tend to co-occur in images and also in text documents, which could give rise to a latent factor related to 'cutlery' in vision and language models alike. It should be expected that the shared features are those that explain a lot of variance in the sensory input, so they will also account for much of the neural-response variance explained by models.

Neural networks are universal function approximators, meaning that they can approximate any specified continuous function over a bounded input domain to any desired degree of accuracy, given a sufficient parameter count. To the extent that diverse systems have ample capacity to capture the statistical structure of the world and ample training data, they may converge towards a platonic representation in which their diverse architectures, objectives and learning rules cease to substantially affect their representations<sup>32</sup>. However, all brains and artificial intelligence models are finite and vastly less complex than the world<sup>36</sup>, so the platonic ideal they are driven towards is bound to remain out of reach. Thus, although models share major variance-explaining features, they also acquire unique features that reflect their distinct architectures, objectives and learning rules, and the corners they must cut to achieve good performance with limited resources. Universal approximation makes models act similarly within a shared training distribution. Yet their hidden unique characteristics lend them distinct inductive biases that can be revealed with out-of-distribution probes.

## Generating stimulus sets that make models disagree

Because different models tend to share features that explain major variance in their shared or overlapping training distributions, studies may fail to detect differences in predictive performance when the models are evaluated on samples from a similar distribution (such as randomly sampled natural stimuli)<sup>17,18</sup>. For model comparison to drive theoretical progress, researchers may therefore need to go beyond the training distribution. Empirical studies using stimuli optimized for model discrimination often reveal considerable performance differences among high-parametric models<sup>29–31,34,37–41</sup>. We refer to such stimuli as *controversial stimuli* because they are controversial among the models: they make the models disagree in their predictions. Like adversarial examples, controversial stimuli result from optimization using an objective defined on the responses of the models. However, adversarial examples aim to make a single model err, whereas controversial stimuli aim to elicit distinct predictions from alternative models. Although this stimulus-generation approach is still new, the early empirical results suggest that optimized stimulus sets can disentangle the predictions even of strong function approximators.

The power to discriminate models depends not only on the experiment but also on the data-analysis technique. When the correspondence between model units and animal neurons or behavioural responses is unclear, encoding models<sup>4,42</sup> or RSA<sup>43,44</sup> can be used to relate the two systems. Both approaches have seen substantial recent advances<sup>45–51</sup>. Encoding models and RSA are closely related and complementary, and together provide a powerful set of tools for model comparison<sup>45,52,53</sup>. The choice of analytic approach has implications for stimulus design, which we discuss in detail below.

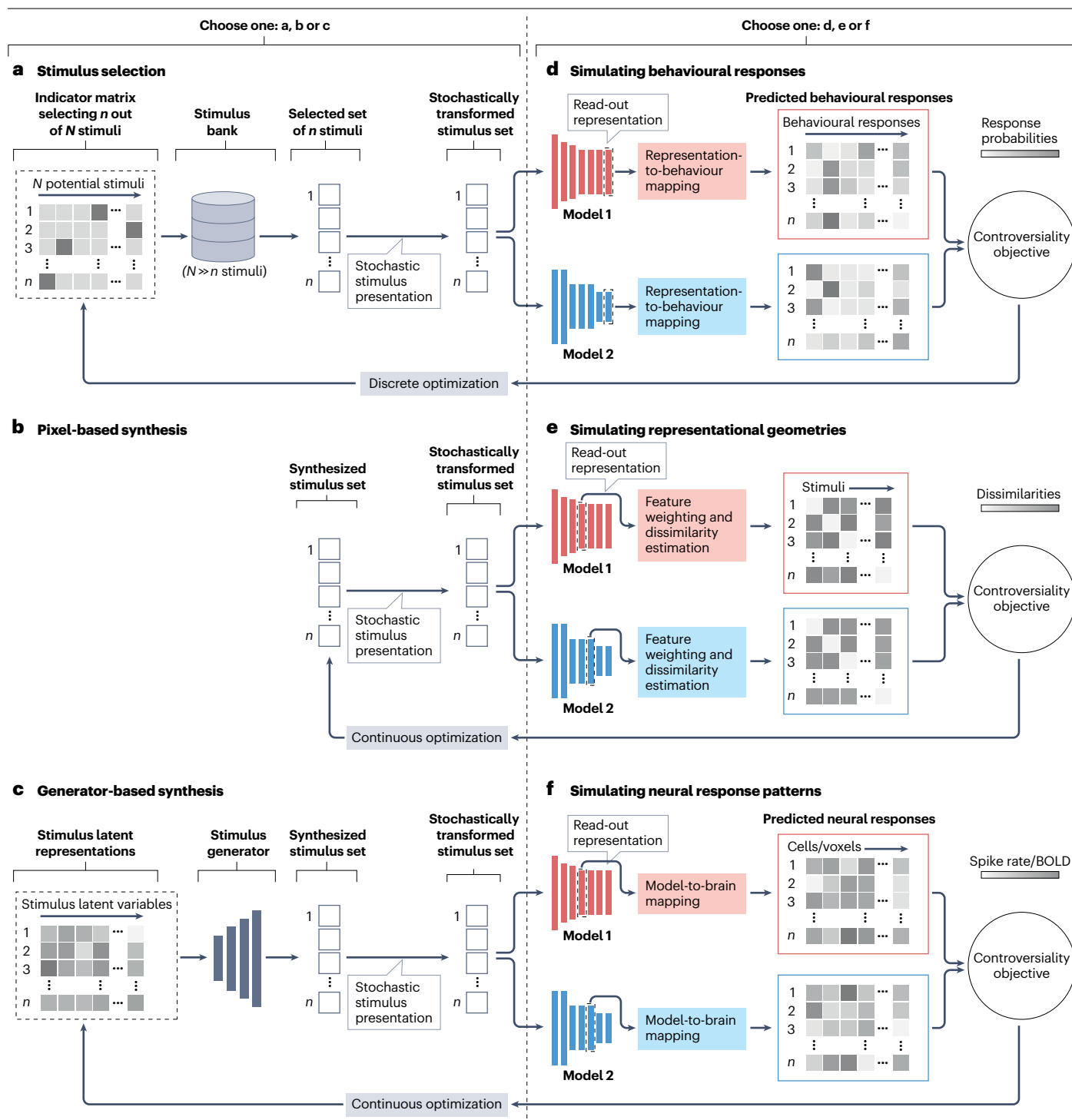
Figure 1 provides an overview of two overarching choices researchers must make: the type of search space in which candidate stimuli are to be selected or synthesized (Fig. 1a–c) and the type of outcome to be measured, which can be behavioural responses, representational geometries or neural responses (Fig. 1d–f).

Given these two choices, selecting or synthesizing stimulus sets that elicit distinct predictions in alternative models is an optimization problem. To optimize the experimental stimuli, researchers need to define four components: the alternative models, each mapping stimuli to predictions about experimental measurements of neural or behavioural responses; an optimization objective that quantifies how controversial a stimulus set is among the models; a stimulus search space that specifies a parameterization for the stimulus or stimulus set to be optimized and the set of admissible stimuli within that space; and an optimization procedure, which enables navigation of the search space towards a controversial stimulus set. We now explain these four components in turn.

## Neural network models predicting neural or behavioural responses

Current work often repurposes neural network models from the machine learning literature and maps their internal representations or outputs to neural or behavioural measurements. We use the term *neural network model* to refer to both repurposed machine learning models and models built or adapted to implement neuroscientific hypotheses. As the field matures, we expect models to gain more neuroscientifically motivated features.

Each neural network model, indexed by  $m$ , predicts responses  $y$  to stimuli  $\xi$ . In the simplest case, each model  $m$  defines a conditional



probability distribution,  $p(y|m, \xi)$ .  $\xi$  specifies the stimulus set and  $y$  represents a specific experimental outcome (a set of neural or behavioural responses). However, predictions are often conditional not only on the identity of the model  $m$  but also on its parameters  $\theta_m$ , defining a probability distribution  $p(y|m, \theta_m, \xi)$ .  $\theta_m$  may include both the neural network weights and the parameters of the mapping from the neural network's hidden-layer activations or outputs to the predicted

neural or behavioural data. For predicting behavioural choices, the outputs of the neural networks can often be mapped directly to behavioural decision probabilities (Fig. 1d) without the introduction of additional parameters. For example, the softmax output (a probability distribution across stimulus classes) of a neural network model trained to classify inputs can be interpreted as predicted probabilities of behavioural responses in a classification task. More generally, linear encoding

**Fig. 1 | Optimizing experiments for model discrimination.** Experimental design requires choosing a type of stimulus search space (panels **a–c**) and a type of outcome to be measured (panels **d–f**) before the finer-grained decisions that enable an experimental stimulus set to be optimized. The choice of stimulus search space must balance model-comparison power and ecological validity (panels **a–c**). The stimuli may undergo stochastic transformations, such as the addition of noise, to simulate realistic perceptual conditions (for example, spatial jitter can simulate imprecise fixation and improve the robustness of model predictions for visual stimuli). Selecting stimuli from a stimulus bank via discrete optimization allows strong control because only stimuli included in the stimulus bank can appear in the experiment (panel **a**). Continuously optimizing visual stimuli in pixel space (or stimuli for other sensory modalities in an analogous low-level description) provides a highly flexible search space because, in principle, any displayable image can be reached, subject only to minimal restrictions such as image resolution and colour space (panel **b**). Generating a stimulus set by optimizing the latent variables of a specific stimulus generator (for example, a generative adversarial network<sup>87</sup>, a diffusion model<sup>93</sup> or a differentiable renderer<sup>38</sup>) provides a compromise between control and flexibility: the search space is continuous but confined to a generator-defined manifold (panel **c**). For a candidate stimulus set, researchers read out each model's intermediate representations or outputs and transform them into predictions of empirical outcomes, such as behavioural responses, representational dissimilarities or

neural responses (panels **d–f**). The nature of the predicted outcomes determines what optimization objectives are appropriate. Each objective provides a different operationalization of controversiality, defining in what sense the outcomes predicted by different models are to diverge (see Boxes 1–3 for mathematical definitions). When behavioural responses are predicted, model outputs or layer representations are mapped to a predicted behavioural response distribution (for example, using a softmax for predicting categorical choices). Model discrimination experiments can be designed by finding stimulus sets that produce dissimilar predicted behavioural response distributions across the alternative models (panel **d**). When representational geometries are predicted, layer representations are summarized by representational dissimilarity matrices, optionally incorporating pre-fitted feature weighting. Model discrimination experiments can be designed by finding stimulus sets that yield dissimilar representational dissimilarity matrices (panel **e**). When neural responses are predicted using linear mappings from the model units, model discrimination experiments can be designed by finding stimulus sets that yield dissimilar predicted neural responses (panel **f**). Importantly, whether behavioural responses, representational geometries or neural responses are being predicted, researchers may need to account for measurement noise and uncertainty about the parameters of the model-to-brain or model-to-behaviour mapping. BOLD, blood-oxygen level-dependent responses.

models and RSA are commonly used to relate model representations to neural or behavioural response patterns elicited by stimuli, even when a direct mapping is unavailable.

When linear encoding models are used, the predicted outcomes  $y$  are the responses of  $C$  individual neurons or voxels (Fig. 1f). Each measured response is predicted separately, and the noise is typically assumed to be independent across stimuli. The linear encoding model specifies the likelihood  $p(y|m, \theta_m, \xi)$ , that is, the probability (or probability density) of the observed response vector  $y$  given the model  $m$ , its parameters  $\theta_m$  and the stimulus set  $\xi$ . The likelihood is usually defined using a Gaussian<sup>42</sup> or Poisson<sup>54</sup> noise model<sup>4</sup>. Typically, the number of stimuli with corresponding neural responses is small relative to the number of weights to be fitted, and the individual neuron or voxel responses are affected by substantial trial-to-trial variability and measurement noise. To mitigate overfitting, regularization is generally required. A popular choice is to use a zero-mean Gaussian prior on the encoding weights (ridge regression). The performance of each linear encoding model at predicting the responses of neurons or voxels is evaluated for an independent set of test stimuli. Generalized linear models can also be fitted to predict behavioural data, for example, using the softmax output to predict probabilities of discrete behavioural choices (see refs. 55–57 for examples).

RSA characterizes the representational geometry of a set of stimuli by the multivariate distances among the stimuli in the neural population response space of each model layer and brain region. The model-predicted experimental outcomes  $\hat{y}$  and the observed outcomes  $y$  are then vectors of representational distances (one distance for each pair of stimuli; Fig. 1e). The response-space distance estimates for all pairs of stimuli are often assembled in a square representational dissimilarity matrix (RDM; where the more general term dissimilarity is chosen to include distance estimators that are not distances in the mathematical sense<sup>53</sup>). Alternatively, the pairwise relationships can be captured using a similarity or kernel matrix. RSA is also applicable to behavioural experiments that yield representational similarity or dissimilarity estimates for pairs of stimuli (such as similarity judgements). The likelihood  $p(y|m, \theta_m, \xi)$ , that is, the

probability density of the observed dissimilarity estimates  $y$  given the model  $m$ , its parameters  $\theta_m$ , and the stimulus pairs  $\xi$ , is typically not explicitly defined (but see refs. 58–60 for probabilistic approaches). Instead, RSA requires that an RDM estimator and an RDM comparator be chosen<sup>50</sup>. The RDM comparator measures the similarity between two neural-population representations (each a model layer or brain region). The RDM estimator and RDM comparator jointly define what aspects of the representational geometry matter in measuring the discrepancy between two representations and under what conditions two representations are considered equivalent<sup>48,61</sup>. For example, the model-predicted and observed vectors of representational dissimilarities  $\hat{y}$  and  $y$  can be compared using the whitened Pearson correlation or the whitened cosine similarity. These whitened RDM comparators account for the correlational structure of the dissimilarity estimates and afford nearly the same statistical power for model comparison as the theoretically optimal likelihood-based approach<sup>45,46</sup>, which requires stronger assumptions. When the RDM estimator is the squared Euclidean distance and the RDM comparator is the whitened cosine similarity, then the comparison is mathematically equivalent<sup>46,49</sup> to the linear centred kernel alignment, a measure for comparing representations that is popular in machine learning<sup>62</sup>. Alternative RDM estimators and RDM comparators differ in their bias and variance at the level of the distance estimates<sup>47,63</sup> and at the level of the comparator outcomes, which reflect the similarity between two representations, such as a model layer and a brain region<sup>64</sup>.

For both linear encoding models and RSA, the parametric flexibility of the mapping from the model units to the neural or behavioural response measurements has to be considered<sup>65</sup>. By default, encoding models require  $P \times C$  weights, one for each combination of the  $P$  model units (predictors) and the  $C$  measured response channels (such as neurons or voxels). The factor  $P$  can be reduced by using a structural prior, for example, in a separable encoding model<sup>51,66</sup>, where the matrix of unit weights is defined as an outer product of a weight vector across feature maps and a weight vector across spatial locations. Whereas linear encoding models still always require

**Table 1 | Three levels of parametric flexibility of neural network models**

| Design consideration                                     | No flexibility                                                                  | Linear flexibility                                                                                                            | Nonlinear flexibility                                                                                                                                                                                      |
|----------------------------------------------------------|---------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fitted parameters                                        | Pre-fitted, fixed parameters                                                    | Linear mapping models still to be adjusted                                                                                    | End-to-end nonlinear models still to be adjusted                                                                                                                                                           |
| Will the neural network model be fitted to the new data? | No, the final parameters are considered known                                   | No, the final parameters are considered known                                                                                 | Yes, the model will be fitted to neural data                                                                                                                                                               |
| Will a linear mapping model be fitted to the new data?   | No, specific feature weighting is assumed                                       | Yes, a linear mapping model will be fitted to neural data                                                                     | Yes, subsumed in the nonlinear model                                                                                                                                                                       |
| Hypothesis to be evaluated                               | Architecture, training objective, learning rule, training data, feature weights | Architecture, training objective, learning rule, training data                                                                | Architecture                                                                                                                                                                                               |
| Application of controversial stimuli                     | Controversial stimuli are straightforward to generate                           | Mapping flexibility needs to be accounted for when generating controversial stimuli and evaluating responses to these stimuli | Controversial stimuli can help reveal whether the architectures are functionally distinct. If not, training to explain responses to the controversial stimuli will efficiently drive models into alignment |

a substantial number of parameters, RSA enables the evaluation of representational predictions without any parameters for a model-to-brain mapping<sup>43,44,47,50</sup>: the RDM for each model layer and brain region is indexed only by the stimuli (containing a dissimilarity for each pair of stimuli). Any two RDMs can thus be directly compared. However, some flexibility is often still desired. For example, one may assume that a model has the right set of features (tuning curves), but that the relative prevalences of those features in the neural population are unknown. Weighted RSA enables modelling the relative prevalence of each feature in the model using one weight per unit (so  $P$  rather than  $P \times C$  weights) or one weight for each set of units (for example, for each feature map or each layer)<sup>47,67–70</sup>. If full linear flexibility is desired, RSA can also be used to evaluate representational geometry predictions of linear encoding models<sup>68</sup>.

In addition to the parameters involved in mapping the model to the empirical predictions, neural network models have many connection-weight parameters, whose values affect model predictions. These parameters can be set by training models to achieve good task performance; alternatively or additionally, models can be fitted to neural and/or behavioural data.

When planning a new experiment, researchers need to consider their uncertainty about the neural network parameters  $\theta_m$ , which include both the weights of the neural network and the parameters defining the mapping from neural network units to experimental outcomes. The degree to which researchers are uncertain about these parameters determines how vaguely or specifically they can define the scientific hypotheses they wish to adjudicate among. More flexible models make less specific predictions and will tend to be harder to distinguish empirically<sup>35</sup>. However, a model that implements one specific version of a hypothesis may fail to fit the data, even if the hypothesis is correct.

If researchers are certain about all parameters before experimental design (a scenario we refer to as ‘no flexibility’), controversial stimuli will be relatively easy to generate, as the predicted responses for each stimulus set are unequivocally defined. However, if researchers are still considering adjusting either the parameters of a linear model-to-brain or model-to-behaviour mapping (‘linear flexibility’) or the parameters of the neural network model itself (‘nonlinear flexibility’), a more complicated approach may be needed. The three levels of remaining model flexibility allowed when analysing the neural or behavioural responses to the controversial stimuli (no flexibility, linear and nonlinear flexibility) are summarized in Table 1. Because the design of controversial stimuli critically depends on the level of flexibility, we now discuss each level in turn.

## No flexibility: all parameters pre-fitted and considered certain.

Pre-fitted models can be obtained by training each neural network model on a task (such as image classification or next-token prediction). If a model-to-brain or a model-to-behaviour mapping is needed (for example, to predict individual responses of neurons or voxels with encoding models or to allow some flexibility in RSA), then this mapping must also be pre-fitted. Once all parameters are fitted, they are considered known and fixed. Each model’s internal representations or outputs can then be compared with neural or behavioural data without further parameter adjustment. When designing controversial stimuli, the likelihood function  $p(y|m, \xi)$  of each pre-fitted model  $m$  is straightforward to evaluate because one need not consider different possibilities for parameters  $\theta_m$ . The likelihood depends only on the model  $m$  and the stimuli  $\xi$  and can be plugged into one of the controversy objectives discussed below (see the ‘Optimization objectives’ section).

Pre-fitted models make very specific predictions. By evaluating such a model, researchers test a computational hypothesis that specifies not only the architecture, training objective, learning rule and training data, but also the relative prevalence in the neural population of each tuning function (and the specific linear encoding, if any). The specificity of their predictions makes pre-fitted models easy to evaluate by experiment. In particular, certainty about the parameters simplifies model comparisons and the optimization of controversial stimuli. Because pre-fitted models do not hedge their bets, but make strong, specific predictions, they are often easy to reject experimentally.

The main limitation of using pre-fitted models is that a model might be rejected on the basis of a premature commitment to particular parameter settings, even when the more general computational hypothesis the model is meant to represent is correct. A model may correctly capture the computations but not the relative prevalence in the neural population of different tuning functions. For example, a qualitatively correct model of V1 that has a quantitatively incorrect receptive-field distribution may fail to accurately predict neural or behavioural responses. Model predictions are often improved substantially when some flexibility is allowed. Even low-parametric flexibility, such as a weighting of the features (reflecting the unknown prevalences of different tuning functions in the neural population), can improve prediction accuracy substantially<sup>17,67–70</sup>. Theoretical progress through model comparison requires that researchers are honest about their uncertainty about the parameters.

**Linear flexibility: model-to-brain or model-to-behaviour mapping considered uncertain.** Researchers usually cannot be certain about the linear mapping from the trained neural network model to the

brain or behavioural data. The parametric flexibility of linear mapping models ranges from a small number of parameters (for example, one weight per layer or feature map in weighted RSA)<sup>67</sup>, through a larger number of parameters in a separable encoding model<sup>51</sup>, to the full  $P$ -unit-by- $C$ -response linear encoding model<sup>4,42,71</sup> (Fig. 1e,f).

Models with linear mapping flexibility make less specific predictions than pre-fitted models. By evaluating such a model, researchers are testing a computational hypothesis that specifies the architecture, training objective, learning rule and training data, but not the specific linear encoding or relative prevalence in the neural population of each tuning function. Generating controversial stimuli for models with flexible model-to-brain or model-to-behaviour mappings is more challenging because the uncertainty about the mapping parameters must be taken into account. Ideally, the stimuli should be controversial for any particular choice of the unknown parameter settings. To account for linear flexibility of the models, the unknown parameters can either be marginalized over (ideally analytically<sup>72</sup>) or chosen such that models are maximally confusable, thus ensuring model discriminability in the worst-case scenario.

**Nonlinear flexibility: weights of the neural network model considered uncertain.** The highest level of model flexibility is to consider all nonlinear parameters of the models uncertain at the time of experimental design (see refs. 73–76 for examples). This approach is sometimes referred to as ‘end-to-end fitting’ (emphasizing that all parameters are being trained) or ‘finetuning’ (when a task-pre-trained model is further trained to fit the experimental data by adjusting its neural network weights). Models with all parameters still to be fitted make the least specific predictions. By evaluating such a model fitted to the new data starting from random weights, researchers are testing a computational hypothesis that specifies only the architecture (for example, testing whether a certain architecture can approximate the neural data well<sup>76</sup>). When the model is first pre-trained on a task and then finetuned, the hypothesis may extend to the task as well.

The greater flexibility of end-to-end fitting on the new data renders the models harder to distinguish, thus creating a greater need for controversial stimuli, but it also makes controversial stimuli harder to generate. Marginalizing over the unknown parameters or finding parameters that render models maximally confusable may be intractable in practice. To meet this challenge, it is attractive to consider iterative experiments with controversial stimuli, where new stimuli are generated in each round to uncover differences among models as the parameters converge to their best settings across experiments.

## Optimization objectives

To pick a stimulus set ( $\mathcal{X}$ ), researchers need to define an objective to be optimized. Intuitively, the objective should ensure that the models make predictions that are as distinct as possible. However, deriving an objective that quantifies this intuitive notion of distinct predictions requires explicitly articulating the underlying goal.

The goal may be to learn which of a number of brain-computational models best accounts for a brain representation, such as the neural population code in a specific cortical area for some domain of stimuli. If researchers assume that their models include the true model, then the question becomes: which is the true model? Given a dataset of outcomes ( $\mathbf{y}$ ) generated by an experiment using the particular stimulus set  $\mathcal{X}$ , this question is elegantly addressed by probabilistic inference about the identity of the true model: researchers calculate the likelihood of each model and use Bayes’ rule to update their prior belief

about which model is the true model. But how should a stimulus set  $\mathcal{X}$  be picked that would best inform this update?

In the Bayesian framework, a natural criterion for evaluating a potential experiment is the expected information gain (EIG)<sup>77–79</sup>. The EIG is the expected reduction of the researcher’s uncertainty about which model is true, where uncertainty is measured as the entropy of the posterior distribution over model identity (Box 1). The EIG can also be understood as the Shannon mutual information  $I(M; Y)$  between the experimental outcome  $Y$  and categorical model identity  $M$  (capital letters signify random variables here). Owing to the symmetry of mutual information, the EIG that the experimental outcome would provide about the identity of the true model is equal to the EIG that an oracle revealing which model is true would provide about the experimental outcome

$$I(M; Y) = H(M) - H(M|Y) = H(Y) - H(Y|M). \quad (1)$$

The experimental outcome is unobserved at the stage of experimental design, so calculating the EIG requires taking the expectation over all possible outcomes. Figure 2 shows an example of estimating and optimizing EIG in a simple setting. When optimizing the expected gain of information about model identity (equation (1)), the aim is to identify the data-generating model, not to most accurately estimate its parameters. Model parameters then are nuisance variables whose influence must be accounted for although they are not the object of the inference. Researchers must therefore marginalize over the model parameters. For example, for linear encoding models, this requires marginalizing over all potential feature weight assignments. Optimizing this criterion will entail a reduction of uncertainty about model parameters only to the extent that reducing parameter uncertainty for the models helps infer which of them is the true model.

The normative Bayesian approach is important as an ideal to aspire to, but it can be difficult to implement in practice. For example, estimating  $H(M|Y)$  requires computing the expectation across all possible experimental outcomes  $\mathbf{y}$  of the entropy of the posterior  $p(M|\mathbf{y})$ . The experimental outcomes  $\mathbf{y}$  predicted by the models could be behavioural responses to different stimuli, responses of individual neurons to different stimuli or representational distances in a neural population for different pairs of stimuli (see the ‘Controversial stimuli in practice’ section for examples of studies using each of these outcome types). For models with many parameters (those with linear or nonlinear flexibility; see the ‘Linear flexibility: model-to-brain or model-to-behaviour mapping considered uncertain’ and ‘Nonlinear flexibility: weights of the neural network model considered uncertain’ sections), inference of  $p(M|\mathbf{y})$  is often already too difficult even for a single, specific experimental outcome  $\mathbf{y}$ , unless the model is chosen wisely to make such inference tractable<sup>72</sup>. Data analyses aiming to compare models with many parameters therefore often use measures of predictive accuracy (for example, a correlation coefficient between empirical and predicted responses) rather than Bayesian inference. Such measures typically rely on maximum-likelihood model parameters and use cross-validation to mitigate bias due to overfitting<sup>47</sup>. These complications motivate consideration of alternative optimization objectives that are computationally tractable and account for the actual inferential procedure that is to be used to analyse the experimental results.

One alternative objective that maintains important elements of the Bayesian framework is to maximize the probability of recovering the true data-generating model (Box 2): the model-recovery accuracy. This method is highly versatile because the inferential procedure used

## Box 1 | Maximizing the experiment's expected information gain

The normative Bayesian perspective on experimental design derives from probabilistic inference and information theory. Say researchers performed an experiment using design  $\xi$  and obtained data  $\mathbf{y}$ . The normative approach to assessing which model is true is to use Bayes' rule to infer the posterior over the model index  $m$

$$p(m|\mathbf{y}, \xi) = \frac{p(\mathbf{y}|m, \xi)p(m)}{\sum_{m'} p(\mathbf{y}|m', \xi)p(m')} \quad (2)$$

where  $p(m)$  is the prior probability over the models. If the models have parameters, the likelihood  $p(\mathbf{y}|m, \xi)$  used in equation (2) can be computed by marginalizing over the model parameters  $\theta_m$

$$p(\mathbf{y}|m, \xi) = \int p(\mathbf{y}|\theta_m, m, \xi) p(\theta_m) d\theta_m. \quad (3)$$

The researchers' goal is to reduce their uncertainty about which model is correct, so the utility of an experiment  $\xi$  is the expected reduction of the entropy of their belief distribution over the models that performing the experiment would entail. This entropy reduction equals the expected information gain (EIG) as researchers update their prior in light of the data  $\mathbf{y}$  to obtain the posterior. Computing the EIG requires marginalizing over all possible true models, their parameters, and all possible experimental datasets  $\mathbf{y}$  they might generate.

Before marginalizing over these variables, it is useful first to consider the information gain that experiment  $\xi$  would provide if it yielded a particular dataset  $\mathbf{y}$ . The information gain is the entropy

reduction from the prior  $p(m)$  to the posterior  $p(m|\mathbf{y}, \xi)$  and serves as the local utility  $u(\xi, \mathbf{y})$ <sup>77</sup>

$$u(\xi, \mathbf{y}) = H(M) - H(M|\mathbf{y}, \xi) = -\sum_m p(m) \log p(m) + \sum_m p(m|\mathbf{y}, \xi) \log p(m|\mathbf{y}, \xi). \quad (4)$$

By averaging the local utility  $u(\xi, \mathbf{y})$  over all possible outcomes, a scalar global utility  $U(\xi)$  for experiment  $\xi$  is obtained

$$U(\xi) = \int p(\mathbf{y}|\xi) u(\xi, \mathbf{y}) d\mathbf{y}. \quad (5)$$

The experimental outcome  $\mathbf{y}$  depends on the experiment  $\xi$ , the data-generating model  $m$  and its parameters  $\theta_m$ . The global utility (the EIG), therefore, is an expectation over possible data-generating models and their parameters

$$U(\xi) = \mathbb{E}_{m, \theta_m, \mathbf{y}} u(\xi, \mathbf{y}) = \sum_m p(m) \int p(\theta_m) \int p(\mathbf{y}|\xi, m, \theta_m) u(\xi, \mathbf{y}) d\mathbf{y} d\theta_m. \quad (6)$$

Note that the local utility defined in equation (4) does not depend on the true model. In the ideal world of Bayesian inference, the reduction in the entropy of the posterior belief over the models is all researchers need to consider: if their inferences give them well-calibrated uncertainty, approaching certainty leads researchers to the correct model. The optimal experiment is  $\xi^* = \arg \max_{\xi} U(\xi)$ .

for comparing models given the experimental outcome is not limited to Bayesian inference. However, to estimate the model-recovery accuracy, researchers must still specify priors for the models and their parameters and marginalize over these unknowns.

A second alternative optimization objective one can adopt, following the frequentist school of thought, is the discrepancy among model-predicted distributions (Box 3). Specifically, if a data-generating model is compared with an alternative model using the two models' log-likelihood ratio, then a natural measure of discrepancy is the Kullback–Leibler (KL) divergence between their predicted distributions<sup>80</sup>. If the alternative model has unknown parameters, these can be adjusted to minimize the chosen discrepancy criterion, so as to consider the worst-case scenario in which the two models are least discriminable. Stimuli are then chosen to maximize this worst-case discrepancy, turning the optimization into a maximin problem (maximization of a minimum value). If the data-generating model has unknown parameters, the expectation of the worst-case discrepancy must be taken with respect to the possible parameter values (via either integration or sampling). As the true model is unknown during experimental design, researchers may sample the model from a uniform prior – assigning each model equal probability – and take the expected value. When comparing each model with multiple alternative models, researchers can either average the divergences across all pairwise comparisons or consider the most challenging comparison (the smallest divergence). Maximizing the discrepancy of the predicted experimental outcome distributions is a highly intuitive objective, and in some cases, it enables researchers to capitalize on analytical solutions to the optimization problem<sup>81–84</sup>.

The three optimization objectives discussed (Boxes 1–3) are distinct in general. An experimental design that maximizes the expected discrepancy between two models' predicted distributions (Box 3) does not necessarily maximize the model-recovery accuracy (Box 2) or the EIG (Box 1).

A simpler and more computationally affordable approach to addressing uncertainty about model parameters is to apply iterative cycles of model fitting, stimulus optimization and experiment. Such an iterative approach has been used successfully in studies manipulating neural responses to test the predictions of a single flexible model<sup>66,73,85</sup>. An iterative approach simplifies the computations by settling for point estimates of the model parameters at each step. The estimates are updated iteratively, such that, ideally, each model converges to its optimal fitted instance. At each iteration, the fitting and evaluation should take advantage of all previously recorded responses to stimuli, which should include natural stimuli to ensure that the models do not drift away and lose ecological validity. Unlike fully Bayesian adaptive experimental design, iterative fitting is not guaranteed to converge to the correct solution. Future work should explore how well such heuristic approaches work in practice.

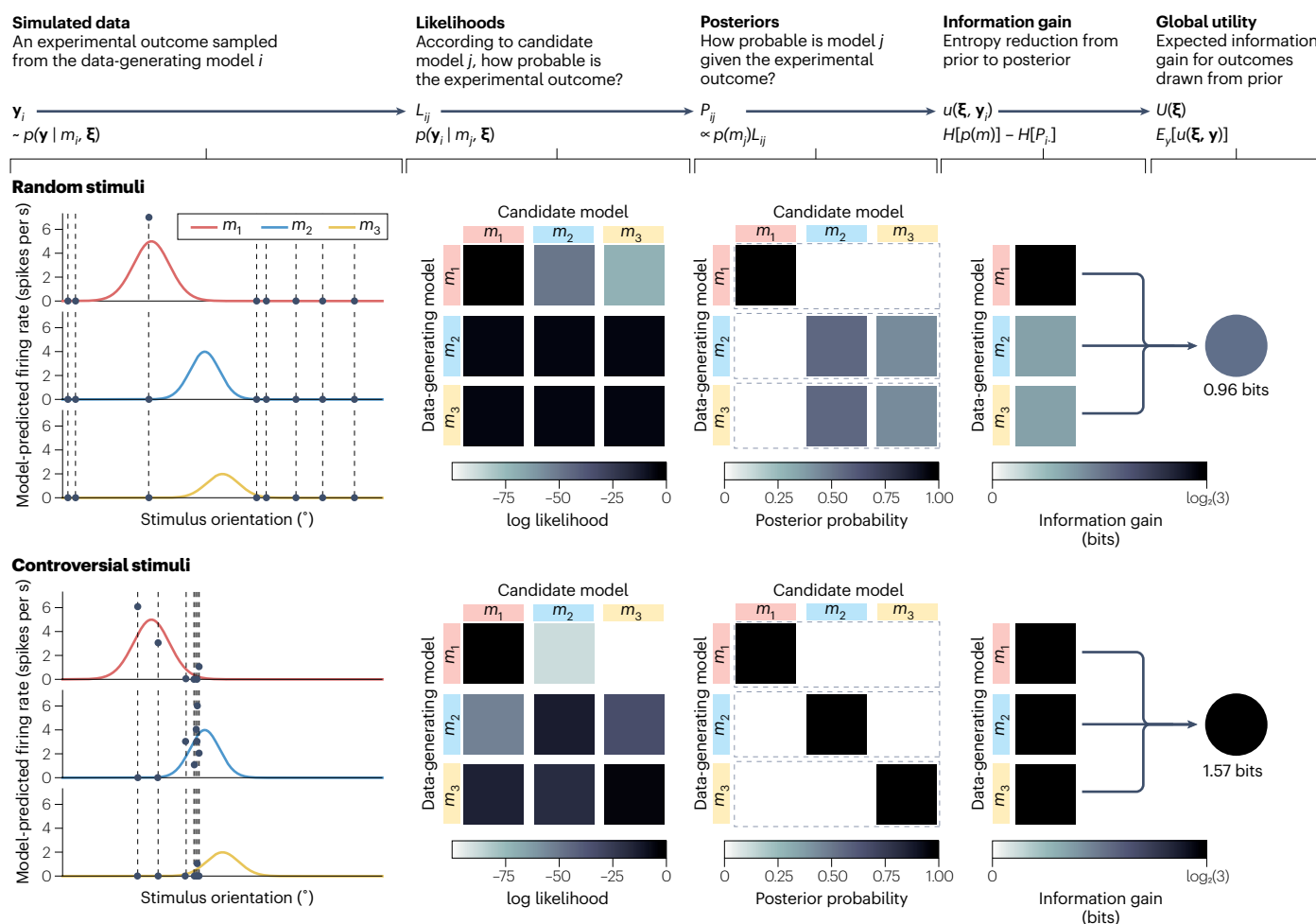
### Stimulus search spaces

Once the objective to be optimized has been defined, the space within which the optimization procedure searches for a stimulus set must be specified. This definition includes the admissible stimuli (the set of potential solutions) and the stimulus parameterization (the optimized variables). To be ecologically valid, neuroscientific computational models should explain brain function under natural conditions, so the

natural distribution of stimuli (those individuals encounter during their daily lives) has a privileged status. One approach is to restrict the experimental stimuli to stimuli that are either natural or naturalistic (not too far from the natural distribution). A further restriction may be desirable for models that have a specific stimulus domain over which they are intended to operate. For example, models of reading or face identification may be intended only to make predictions for text or face images, respectively. However, to the extent that researchers take a

model seriously as a mechanistic model, it should also make valid predictions when challenged with stimuli outside its designated stimulus domain.

A simple way to define the admissible stimuli is as a pre-defined collection of stimuli, which could be restricted to those relevant for the model (Fig. 1a). The experiment can then be optimized by selecting a subset of the admissible stimulus set that maximizes a model-discrimination objective<sup>30,31,37,39,86</sup>. This approach can be



**Fig. 2 | Estimating and optimizing expected information gain using a Monte Carlo approach.** In this simulation, we compare the predictions elicited by two sets of eight stimulus orientations. For each set, the predictions of three alternative toy models of an orientation-selective neuron ( $m_1$ ,  $m_2$  and  $m_3$ ) are depicted in different colours (red, blue and yellow). In the top panel, the orientations are randomly sampled, and in the bottom panel, the orientations are optimized such that empirically measured responses to these stimuli will enable us to adjudicate among the three data-generating models. Each toy model predicts Poisson-distributed spike counts with a mean firing rate parameterized by a scaled von Mises tuning curve over stimulus orientation (left column). The outcome of a single simulation for each data-generating model is plotted (black dots showing firing rates) for each stimulus set ( $\xi$ , dashed grey lines). For each simulation from a data-generating model, a log-likelihood value is estimated for each candidate model using the natural logarithm (second column). Applying Bayes' rule yields posterior probabilities over candidate models for each of the three simulations (third column). The information gain the stimulus set achieves in a particular simulation

is the reduction in entropy of the posterior relative to a prior belief over the three models (here, assigning equal prior probability to each model). This information gain serves as a local utility, specific to each simulated outcome (fourth column). Taking an expectation over the three data-generating models and over many simulated outcomes yields a global utility: the expected information gain afforded by the stimulus set (Box 1). Here, the utility was estimated by averaging across 3,000 simulations. The random stimulus set was drawn from a uniform distribution over the interval  $[0^\circ, 180^\circ]$ . This set yields simulated spike counts that do not discriminate between  $m_2$  and  $m_3$ . By contrast, maximizing the expected information gain by differential evolution yields a stimulus set concentrated at orientations where the three toy models disagree in their predicted firing rates. Model-recovery accuracy was near perfect for the controversial stimulus set (99.83%) but substantially lower for the randomly sampled stimulus set (73.80%). Similar estimation and optimization procedures can be applied to more elaborate stimulus parameterizations, enabling the selection or synthesis of complex stimuli to support adjudication among neural network models.

## Box 2 | Maximizing the experiment's expected model-recovery accuracy

A natural alternative to maximizing expected information gain (EIG) is to maximize the probability of selecting the true data-generating model. This is equivalent to maximizing the expected model-recovery accuracy (see ref. 126 for example). Here, the local utility is defined as success in correctly identifying the data-generating model given an experimental design  $\xi$ , ground-truth model  $m$ , experimental outcome  $y$  and a model comparison procedure. Specifically, if the fit of each model to the simulated data  $y$  is quantified by a score  $S(y, \xi, m)$ , the local utility can be defined as

$$u(\xi, y, m) = \mathbb{I}[\forall m' \neq m : S(y, \xi, m) > S(y, \xi, m')] \quad (7)$$

Equation (7) will equal 1 if the model with the highest score is the true model, and 0 otherwise. If the model set includes a model that is not distinguishable from the true model, equation (7) will always return 0.

A more permissive criterion is the fraction of alternative models that the true model dominates, which is the average of the model-recovery accuracy over all pairwise model comparisons

$$u(\xi, y, m) = \frac{1}{|\mathcal{M}| - 1} \sum_{m' \neq m} \mathbb{I}[S(y, \xi, m) > S(y, \xi, m')] \quad (8)$$

The global utility is the expected value of the local utility (equation (7) or (8)) over all possible data-generating models, their parameters and the resulting experimental outcomes

$$U(\xi) = \sum_m p(m) \int p(\theta_m) \int p(y|\xi, m, \theta_m) u(\xi, y, m) dy d\theta_m. \quad (9)$$

Note that this criterion is considerably more versatile than the EIG (Box 1). The model performance score  $S$  can reflect posterior model probabilities, but alternatively it can also reflect non-Bayesian measures such as the goodness or predictive performance of maximum likelihood fits, or even non-probabilistic measures such as the rank prediction accuracy (Spearman or Kendall rank correlation coefficient) or a cross-validated measure of generalization performance. As for the EIG (Box 1), the optimal experiment is  $\xi^* = \arg \max_{\xi} U(\xi)$ .

interpreted as using a discrete parameterization in which each stimulus in the pre-defined collection is represented by a binary parameter that indicates whether it is included in the experimental stimulus set.

Instead of selecting stimuli among a pre-defined collection, the optimization process can select stimuli from a richer discrete or continuous parameter space. This approach is sometimes called 'stimulus synthesis' because the stimulus space contains novel possibilities. Such larger stimulus spaces offer greater flexibility than choice among a pre-defined collection. However, researchers need to consider whether the resulting stimuli are sufficiently ecologically valid.

A key choice in stimulus synthesis is the parameterization of the stimuli. For images, one simple parameterization is pixel intensities (Fig. 1b). When optimizing pixel intensities, the admissible stimuli may include all displayable images. Stimulus synthesis can be guided more strongly by using a different parameterization, such as the latent representation of a stimulus-generating model that defines a manifold of naturalistic stimuli (Fig. 1c). For example, researchers can search for controversial stimuli within the latent representation of a neural network-based generator, such as a generative adversarial network trained to generate naturalistic images<sup>87</sup>. Such an approach can be understood as assuming a naturalistic prior that excludes many possible stimuli. Consider the manifold of possible images spanned by a model trained to transform a vector of 200 numbers into an image. The resulting space of images is drastically smaller than the space of all possible images (Fig. 12.20 in ref. 88).

If the models have a limited scope, such as faces or text, the admissible stimuli can be restricted by using a domain-specific generator. For example, the latent face space of the Basel Face Model<sup>89</sup> is a continuous space of 3D shapes and textures of human faces<sup>38,90</sup>. Recent advances in generative artificial intelligence, including generative adversarial networks as well as diffusion and flow-based models<sup>91–93</sup>, enable researchers to design highly naturalistic controversial stimulus sets by continuous optimization. However, using a generator does not guarantee ecological validity. In contrast to randomly sampling stimulus latent representations, optimizing them can strain the generator and yield stimuli that lie outside the intended naturalistic manifold.

The optimized latent representations are then like adversarial examples for the generator, making it fail at its task of generating natural stimuli. Additional safeguards – such as manual or automatic inspection – may therefore be necessary.

The set of admissible stimuli can be restricted either by using a pre-defined collection of stimuli or by using a parametric stimulus space that does not allow for all possible stimuli. Either way, restricting the stimulus set imposes a hard prior: a constraint that assigns zero probability to certain stimuli (typically, the vast majority of them), thus excluding them from consideration. A parameterization can additionally introduce an implicit soft prior (a bias rather than a strict constraint) by shaping the optimization trajectory such that some stimuli are more likely to be included than others. When using local optimization procedures such as gradient descent to find local maxima of controversiality in the search space of stimulus sets, solutions that are closer to the initialization values of the stimulus parameters are more likely to be selected than solutions that are further away. Even among parameterizations that yield the same set of admissible stimuli, the choice of parameterization matters because it determines the distance of different solutions from the initialization and the gradients that shape the trajectory in the search space. An invertible linear transform of a given parameterization, for example, can change which of multiple local optima the search converges to, thereby changing the outcome of optimization<sup>94</sup>.

Stimulus synthesis can be further regularized by introducing an explicit soft prior (expressed as a penalty term) to the optimization objective (for example, ref. 66 used a total variation penalty to promote spatially smooth images). The process can also be regularized by inserting a stochastic stimulus-presentation model into the stimulus-optimization pipeline (Fig. 1a–c). A stochastic stimulus-presentation model can help create stimulus sets that are robustly controversial. For example, if images are jittered (slightly shifted at random) during optimization<sup>94</sup>, they will elicit contrasting responses from the models irrespective of their precise placement in the input array, accounting for random variation in stimulus presentation during an experiment in which an animal may not fixate perfectly.

Penalty terms and stochastic variation are both forms of regularization that change the loss landscape traversed by the optimization process, thus altering the final solution.

## Optimization procedures

We have discussed the choices of a stimulus search space and an optimization objective. A separate decision is what algorithm to use to optimize the objective. For any given objective, a variety of algorithms, called optimization procedures, could be used to find a controversial stimulus set. As neural network models are complex and nonlinear, it will usually be impossible to find a set of controversial stimuli that is guaranteed to be globally optimal. However, it is often possible to find stimulus sets that are far more controversial than sets of randomly sampled stimuli and yield substantial increases in power for inferential model comparison. In this respect, the problem of optimizing a stimulus set for controversiality resembles the problem of optimizing a neural network model for task performance: in both cases, iterative local improvements can yield powerful solutions, although those solutions are not guaranteed to be globally optimal.

A central distinction for the choice of optimization procedure is whether the search space of candidate stimulus sets is continuous or discrete. Continuous and discrete optimization are similar in some

respects. Even if the search space is discrete and finite, researchers will usually not be able to evaluate the optimization objective for each possible stimulus set and thus will still have to rely on iterative search. The continuous and discrete scenarios are similar in that the optimization procedure will typically find a solution by taking small steps through a stimulus parameter space, with each step increasing the controversiality. Once the process has converged on a local optimum, it will not be possible to find further improvements. However, continuous and discrete scenarios differ substantially in terms of the nature of the steps taken in the stimulus search space. In a continuous search space of stimulus sets, arbitrarily small steps can be taken and, if the controversiality objective is differentiable, its gradient indicates the direction of steepest local increase in controversiality. By contrast, in a discrete space a step may correspond to exchanging a stimulus for a different stimulus, or to a discrete modification of one of the stimuli. The optimization procedures required when the search space of stimulus sets is continuous or discrete therefore fall in different classes.

For discrete stimulus parameterizations, even a simple greedy search, such as starting with a random stimulus set of the desired size and randomly perturbing the stimulus set to improve the controversiality objective, can be effective. For the stimulus-selection problem, for example, in which an experimental stimulus set is to be assembled from

## Box 3 | Maximizing Kullback–Leibler divergences between model-predicted distributions

The frequentist perspective on optimal experimental design suggests directly maximizing the discrepancy between the predictions of different models. Assume that one compares two models,  $m_1$  and  $m_2$ , and that, without loss of generality,  $m_1$  denotes the true model and  $\theta_1$  denotes its true parameter vector. In this setting, the Kullback–Leibler (KL) optimality criterion is defined as the KL divergence ( $D_{KL}$ ) between the predictions of the two models, with the parameter vector  $\theta_2$  chosen to minimize the divergence<sup>127</sup>

$$I_{2,1}(\xi, \theta_1) = \min_{\theta_2} D_{KL}[p(Y|\xi, m_1, \theta_1) \| p(Y|\xi, m_2, \theta_2)] \\ = \min_{\theta_2} \int_{y \in \mathcal{Y}} p(y|\xi, m_1, \theta_1) \log \frac{p(y|\xi, m_1, \theta_1)}{p(y|\xi, m_2, \theta_2)} dy. \quad (10)$$

The KL optimality criterion can be viewed as the expected log-likelihood ratio between the data-generating model and the alternative model, considering the worst-case setting of the alternative model's parameters (those that render the alternative model most confusable with the true model). Choosing stimuli that maximize this criterion makes the models disagree even under the parameter settings that render them most confusable.

The KL optimality criterion greatly simplifies if one assumes that the models share a multivariate Gaussian error distribution, that is,  $y_1 = \eta_1(\xi, \theta_1) + \epsilon_1$  and  $y_2 = \eta_2(\xi, \theta_2) + \epsilon_2$ , where  $\eta_m$  is a function that maps an experimental design and model parameters to the predictions of model  $m$ , and  $\epsilon_1$  and  $\epsilon_2$  are vectors of normally distributed residuals,  $\epsilon_1, \epsilon_2 \sim N(\mathbf{0}, \Sigma)$ . In this case, the KL optimality criterion reduces to the minimal squared Mahalanobis distance between the two vectors

$$I_{2,1}(\xi, \theta_1) = \frac{1}{2} \min_{\theta_2} [(\eta_1(\xi, \theta_1) - \eta_2(\xi, \theta_2))^T \Sigma^{-1} (\eta_1(\xi, \theta_1) - \eta_2(\xi, \theta_2))]. \quad (11)$$

Note that the covariance matrix  $\Sigma$  may depend on the experimental design  $\xi$ . If one assumes that  $\Sigma$  is a scaled identity matrix  $\sigma^2 \mathbf{I}$  and that  $\sigma^2$  is independent of the experimental design  $\xi$ , then the KL optimality criterion is proportional to the classic T-optimality criterion<sup>81,82,127</sup>

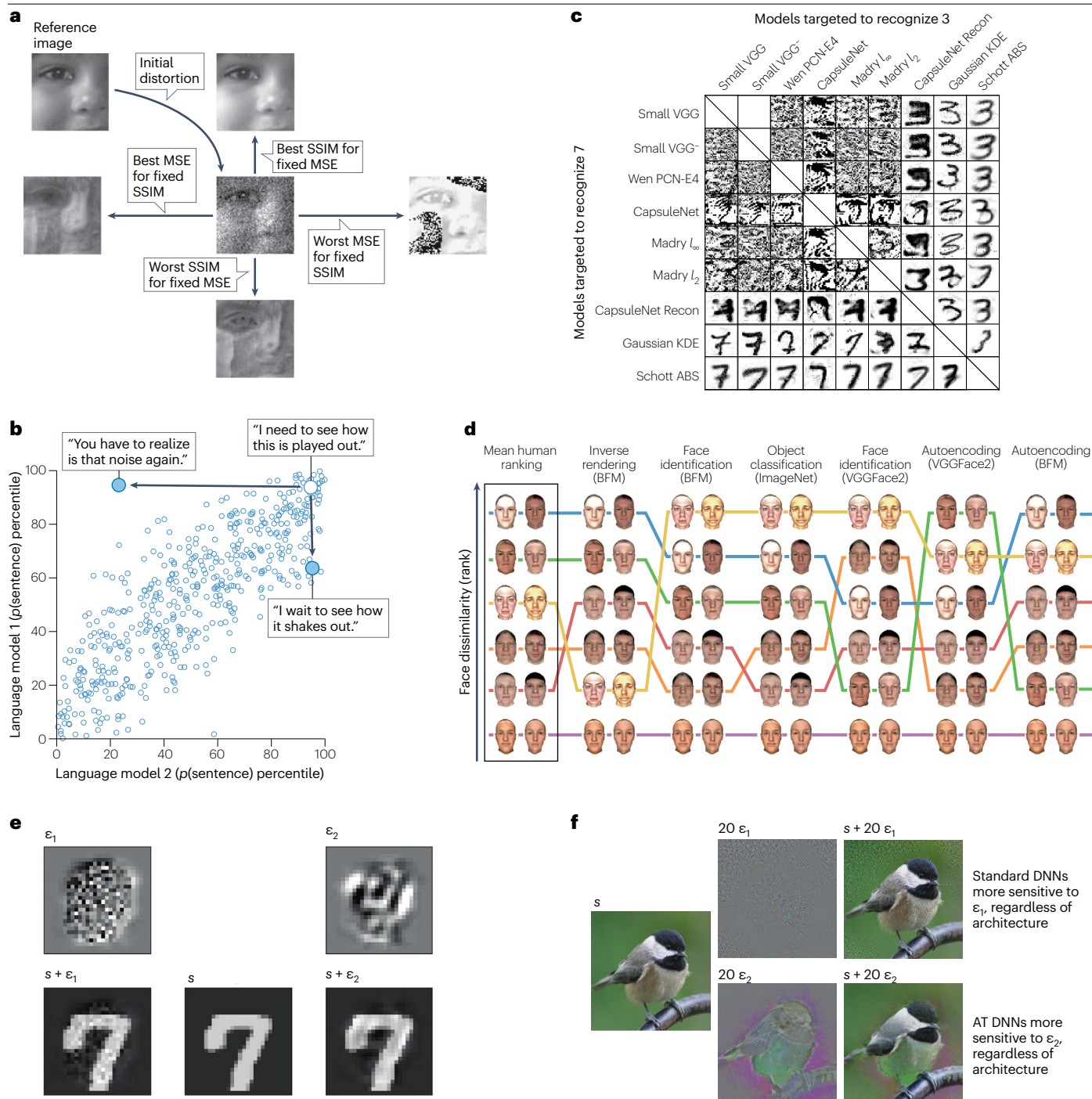
$$T_{2,1}(\xi, \theta_1) = \min_{\theta_2} \|\eta_1(\xi, \theta_1) - \eta_2(\xi, \theta_2)\|^2. \quad (12)$$

In the frequentist perspective, the natural way to generalize the objective specified in equation (12) to the more realistic case where the true model and its parameters are unknown is to take the minimum over these choices

$$U(\xi) = \min_m \min_{\theta_m} \min_{m' \neq m} I_{m',m}(\xi, \theta_m) \\ = \min_{m, m', \theta_m, \theta_{m'}} D_{KL}[p(Y|\xi, m, \theta_m) \| p(Y|\xi, m', \theta_{m'})]. \quad (13)$$

This objective characterizes the worst-case scenario for model discrimination. Note the analogy to equation (6), where models and parameter settings are instead weighted by their probabilities and averaged over. If researchers have trustworthy estimates of the prior over the models ( $p(m)$ ) and the prior over the parameter settings ( $p(\theta_m)$ ) for each model, they can take expectations instead of considering the worst case. Researchers can also interpolate between averaging and minimization by choosing which variables to average over and which ones to take the minimum over. If researchers do not have trustworthy priors over models, they can use an unweighted average over all pairwise model comparisons as in equation (8). All combinations of averaging and minimization yield sensible optimization objectives. As for the other objectives (Boxes 1 and 2), the optimal experiment is  $\xi^* = \operatorname{argmax}_{\xi} U(\xi)$ .

# Review article



a large collection of candidate stimuli, the perturbation can consist in a random replacement of an individual stimulus. The algorithm considers many random replacements and executes the ones that improve the controversy objective<sup>30,34,39</sup>. If the optimization objective is noisy (for example, because its estimation involves Monte Carlo sampling), more principled stochastic optimization approaches<sup>95</sup> or evolutionary algorithms<sup>96</sup> may yield better results.

For continuous stimulus parameterizations, gradient descent and its variants can be highly effective for both deterministic and stochastic

objectives. However, gradient descent requires that the optimization objective be differentiable with respect to the stimulus parameters, so all intervening components, including all models and the stimulus generator (if any), must be differentiable. If the gradient is unavailable because one or more components are non-differentiable, one can use a numerical estimate of the local gradient based on taking small steps around the current stimulus parameters. This approach may not work well when the evaluation of the optimization objective is noisy, because the effect of the small steps can become undetectable under

**Fig. 3 | Stimulus synthesis for model discrimination.** Several studies have synthesized controversial stimuli in different search spaces, including pixel space, generator-defined latent spaces and the discrete space of possible sentences, and used different optimization procedures and controversy objectives. **a**, Maximum differentiation competition<sup>29</sup>. A reference natural image is subjected to an initial distortion and then optimized in pixel space to generate a pair of images to which one perceptual model assigns maximally different distortion strengths ('Best' and 'Worst'), while the predictions of the other model are held fixed. When the distortion strength predicted by structural similarity (SSIM) is manipulated while the distortion strength predicted by mean-squared-error (MSE) is fixed (vertical arrows), human judgements of the perceived quality of the resulting images change more strongly than in the converse scenario (horizontal arrows), supporting the SSIM model. **b**, Discrete optimization. Natural sentences drawn from a corpus (open dots) are considered similarly probable by two language models. Iteratively replacing words yields controversial synthetic sentences (two filled blue dots) that elicit contrasting probabilities, thereby discriminating the two models<sup>39</sup>. **c**, Controversial stimuli for models of classification<sup>37</sup>. A random noise image is optimized in pixel space to elicit distinct classifications in pairs of models. The row label specifies the model that recognizes the digit '7' in the image; the column label specifies the model that recognizes the digit '3'. The labels assigned by models with generative components (last three rows and columns), especially the Schott analysis-by-synthesis (ABS) model, are better aligned with human perception than those assigned by the six discriminative classifiers (first six rows and columns). **d**, Restricting the admissible stimuli for synthesis. Sets of face pairs

are synthesized within the search space defined by a Basel Face Model (BFM) generator to elicit distinct face-dissimilarity ranking predictions from alternative deep neural networks (DNNs)<sup>38</sup>. The left-most column depicts the mean human ranking of one set of face pairs by perceptual dissimilarity. The columns to the right depict the rankings predicted by six different DNNs presented with the same face pairs. **e**, Pixel-space perturbations of a reference stimulus  $s$  (the digit 7, centre image) in directions in which the models have maximally divergent sensitivities. The perturbation  $\epsilon_1$  (top-left) maximally distorts the stimulus representation in an  $L_1$ -regularized autoencoder relative to the distortion in an  $L_2$ -regularized autoencoder; the perturbation  $\epsilon_2$  (top-right) does the converse. The resulting stimuli  $s + \epsilon_1$  (bottom-left) and  $s + \epsilon_2$  (bottom-right) demonstrate the differential sensitivities of the two autoencoders<sup>40</sup>. **f**, The concept of 'principal distortions' generalizes this approach to more than two models. Perturbations of a stimulus  $s$  are shown here scaled by a factor of 20 ( $20\epsilon_1$  and  $20\epsilon_2$ ). The two perturbations maximize variability in the sensitivity among a set of DNNs including both standard and adversarially trained (AT) models<sup>41</sup>. The two images on the right show the original image with each distortion added. AT models were more sensitive to smooth perturbations of colour and brightness around object contours. Panel **a** adapted with permission from ref. 29, ARVO. Panel **b** reproduced from ref. 39, Springer Nature Ltd. Panel **c** reproduced with permission from ref. 37, PNAS. Panel **d** reproduced from ref. 38, CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>). Panel **e** adapted with permission from ref. 40, Simoncelli et al. Panel **f** adapted from ref. 41, CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>).

the noise. A versatile and derivative-free alternative for navigating both continuous and discrete stimulus spaces is evolutionary optimization (as applied in refs. 31,87).

Because the optimization techniques we discuss here are local, the initial stimulus set, the stimulus parameterization and the chosen particular optimization procedure can all influence the outcome. Approximate optimization procedures are not neutral or prior-free methods for choosing the best stimulus set; they impose their own biases, which the researcher should be aware of. Although beyond the scope of this Review, recent advances in the Bayesian experimental design literature offer efficient new ways to optimize experiments (for a review, see ref. 97). In brief, these advances include unbiased Monte Carlo estimators of the EIG<sup>98</sup> (equation (1) and Box 1); variational bounds that can sidestep the computational costs associated with estimating and optimizing this utility<sup>99</sup>; and even amortization of the optimization with a separate neural network model that designs the next trial given previous data<sup>100</sup>. These methods have yet to be integrated into stimulus selection and synthesis for comparisons among neural network models of animal brains, which poses the additional challenge that the search space of stimulus sets is typically high-dimensional.

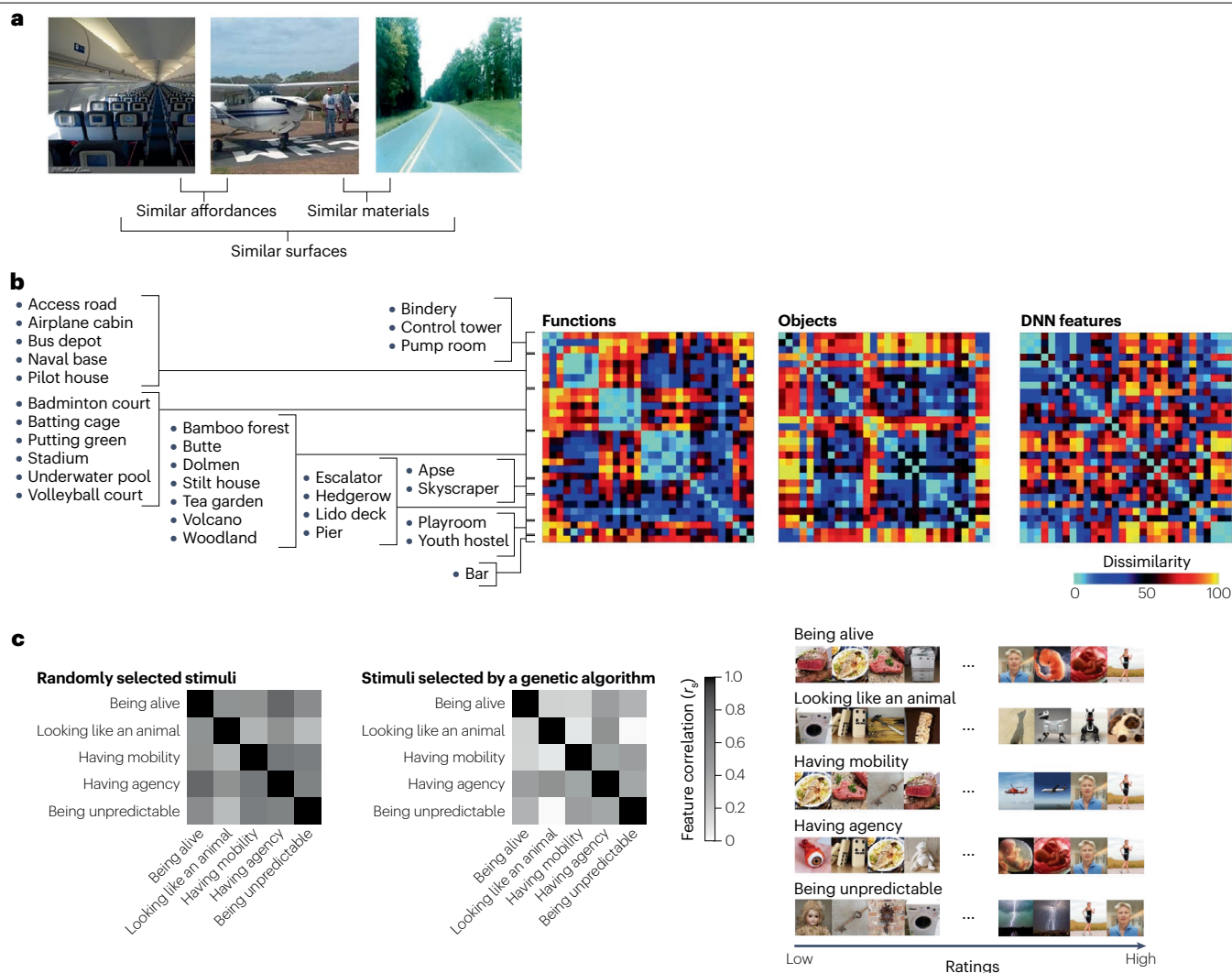
## Controversial stimuli in practice

Now that we have discussed the four requirements for optimizing stimuli for model comparison (the models, the optimization objective, the stimulus search space and the optimization procedure), we review studies that have used stimulus sets designed to be controversial among the models to empirically compare models. Each study made particular choices for the four requirements so as to maximize model disagreement in terms of predictions of the experimental data. The stimuli in these studies are either synthetic (Fig. 3) or selected from a pre-defined collection of admissible stimuli (Fig. 4).

Related methods that are beyond the scope of this Review are low-dimensional stimulus-set optimization for model discrimination (such as optimizing the brightness or timing, a setting thoroughly addressed by the literature on optimal experimental design and adaptive psychophysics; see ref. 101 for example) and high-dimensional stimulus-set optimization with other objectives that do not measure disagreement between models. The latter class includes four approaches for generating stimuli that are optimized for a single model to predict (1) extreme responses<sup>66,73,87,90,102</sup>, (2) responses that are very distinct from those elicited by close-by reference stimuli<sup>85,103–105</sup>, (3) responses that are identical to those elicited by reference stimuli (such that the synthesized and reference stimuli are metamers)<sup>106,107</sup>, or (4) visual illusions<sup>108,109</sup>. Like controversial stimuli, these stimuli are designed to elicit bold model-based predictions. However, in contrast to controversial stimuli, the stimuli are based on a single model, rather than designed to make multiple models disagree in their predictions about neural or behavioural responses.

For practical reasons, studies using optimization of high-dimensional stimuli for model discrimination have relied on heuristics that do not exactly implement the normative objectives (Boxes 1–3). However, these normative objectives are our north star and help to clarify the assumptions and approximations implicit to the heuristic approaches that have been used in practice.

Across domains, the studies reviewed below illustrate a common practical benefit: stimulus sets optimized for model discrimination target conditions under which candidate models make divergent predictions, thereby increasing empirical discriminability relative to conventional stimulus sampling. In some cases, the resulting data support decisive comparisons among models, revealing clear winners; in others, they expose hidden shortcomings of all models. Either way, controversial stimuli provide more informative empirical tests that can reduce ambiguity when multiple high-capacity models achieve similarly high scores on naturalistic test sets.



**Fig. 4 | Stimulus selection for model discrimination.** Several studies have selected stimulus sets from a stimulus bank so that the resulting stimulus set was controversial among alternative models. **a**, A triplet of natural-scene images selected such that each image is the odd one out according to a different image property<sup>86</sup>. **b**, Representational dissimilarity matrices of natural-scene images (descriptions of scenes on the left) selected to decorrelate the predictions of three models: human judgements of functions (whether certain actions can be performed in a scene), human judgements of object categories (whether objects are present in those images) and the penultimate-layer activations of a classification-trained deep neural network (DNN) in response to the same

images<sup>30</sup>. **c**, Spearman-correlation-coefficient matrices for five subjective measures of object animacy, depicting correlations for randomly sampled stimuli, and for a stimulus set selected to decorrelate the five measures by a genetic algorithm<sup>31</sup>. The right panel shows example controversial stimuli ordered by their rank on each of the subjectively rated measures from the prediction-decorrelating stimulus set selected. Panel **a** reproduced with permission from ref. 86, ARVO. Panel **b** reproduced from ref. 30, CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>). Panel **c** reproduced from ref. 31, CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>).

## Discriminating models with *k*-AFC judgements

An effective behavioural paradigm for collecting judgements unaffected by subjective thresholds chosen by participants is the *k*-alternative forced choice (*k*-AFC) task. Consider the problem of selecting a model for predicting which of two corrupted images will be judged as having a higher image quality. Such models can help to assess alternative methods for lossy image compression (methods that reduce file size at the cost of image fidelity). Wang and Simoncelli<sup>29</sup> developed an efficient method called maximum differentiation (MAD) competition for distinguishing two models of perceived image quality (Fig. 3a). They

considered the mean-squared-error (MSE) model and the structural similarity model (SSIM)<sup>110</sup>, which provide two ways of measuring the dissimilarity between the original and the corrupted image. To compare the models, MAD generated two kinds of corrupted-image pairs. In some pairs, the MSE matched between the two corrupted images, whereas the SSIM differed maximally. In other pairs, the SSIM matched, and the MSE differed maximally. For each such corrupted-image pair, participants judged which of the two images appeared less corrupted. The SSIM model predicted these perceptual choices much better than the MSE model<sup>29</sup>. The MAD method can drive high-dimensional stimulus

optimization, enabling efficient comparison of models of perceptual judgements or neural responses and can be extended to compare multiple models in a pairwise fashion<sup>11</sup>. MAD stimuli are not necessarily optimal according to any of the normative objectives for model discrimination (Boxes 1–3). However, holding one model's prediction constant while manipulating the other's prediction provides a highly interpretable experimental design, in which individual comparative judgements can reveal differences between the models.

The restriction of holding the prediction of one model fixed between the two stimuli can be relaxed to an inequality constraint: when modifying a stimulus to reduce the predicted quantity according to one model, the prediction of the other model can be constrained to remain at or above its original level. This relaxation allows more options, which may be needed for the optimization to work. For example, when the stimulus search space is discrete, the optimization procedure may fail to find any stimulus modification that keeps one model's predictions exactly constant. Golan and colleagues<sup>39</sup> used an inequality constraint to compare language-model predictions of human judgements (Fig. 3b). Participants judged which of two English sentences is more likely to be encountered (a linguistic acceptability judgement). Language models estimate how likely a sentence is to occur by assigning it a probability. For each pair of candidate sentences, the model's behavioural prediction was the sentence the model assigned the higher probability to. For pairs of random natural sentences, all models performed indistinguishably well. The researchers generated pairs of synthetic sentences by drawing a natural sentence from a corpus and iteratively replacing words, so as to decrease the probability of the sentence according to one model while keeping it at least equal according to the other model, and vice versa. For the resulting controversial sentence pairs, a generative pre-trained transformer model (GPT-2) better predicted the corresponding human judgements than recurrent neural networks, discriminative language models and simple *n*-gram models.

The studies above used two-alternative forced choice (2AFC) tasks. Controversial stimuli can also be created for forced-choice tasks with more than two alternatives. Greene and Hansen<sup>86</sup> (Fig. 4a) used a triplet-odd-one-out task (an example of a 3AFC task). They constructed 100 triplets of natural-scene images such that, for each triplet, each of the three images is the odd one out according to one of three models. The models were alternative hypotheses about the attributes driving the human judgements, positing that participants will judge which image is the odd one by materials (such as wood), affordances (such as being related to eating) or surfaces (including surface appearance and geometry). They found that affordances most strongly guided human odd-one-out judgements.

## Discriminating models with behavioural judgements of single stimuli

In the studies above, controversial stimuli discriminated among models that predicted forced choices among stimulus pairs or triplets (picking the higher-quality image, the sentence more likely to be encountered or the odd image). An alternative approach is to have participants judge each stimulus independently (for example, estimating image quality or sentence probability on a continuous scale, or assigning a categorical label to each stimulus). Judgements of individual stimuli have been successfully used to compare image-classification models.

Consider the task of classifying an image as depicting a cat or a dog. Both models may recognize cats and dogs well, and both may predict human classifications accurately for natural images of cats and dogs. A researcher can select or synthesize a stimulus that is controversial in

that it is classified as a dog by one model and as a cat by another model. If the models are neural networks whose perceptual mechanisms are not known a priori, the controversial stimulus may contain clues about the different mechanisms used by the two models. Let us assume one model discriminates cats and dogs by their ears and the other by their mouths. In that case, the controversial stimulus might show a creature with a dog's mouth and a cat's ears, or vice versa – revealing the models' contrasting perceptual mechanisms. Most importantly, a human judgement that the controversial stimulus shows either a cat or a dog is guaranteed to provide evidence against one of the models. If the human observer does not perceive either a cat or a dog, then the judgement provides evidence against both models.

Golan and colleagues<sup>37</sup> showed that such individually controversial stimuli provide an effective tool for distinguishing the distinct perceptual mechanisms used by alternative image-classification models. They did so by creating controversial images for each combination of a pair of models and a pair of image labels (such as the digits 3 and 7 in Fig. 3c). Models that incorporated image-generative components predicted human classifications of such controversial stimuli better than purely discriminative classification models. This finding held not only for models trained to classify handwritten digits but also for models trained to classify small images of natural objects.

The method of controversial stimuli as used by Golan and colleagues<sup>37</sup> shares some important features with MAD. Like MAD, it does not maximize the EIG, the model-recovery accuracy or the KL divergence of the predictive distributions (Boxes 1–3). Like MAD, it considers pairs of models and renders the human responses elicited by the stimuli highly interpretable.

Future studies discriminating models with behavioural judgements of singly controversial stimuli may use alternatives to the pairwise approach to comparing multiple models. For example, one could generate stimulus sets that jointly maximize EIG or model-recovery accuracy, taking into account the entire set of candidate models. Furthermore, rather than relying on all class pairs (all pairs of object categories or digits in ref. 37), the choice of class pairs to leverage for model comparison could be left up to the optimization algorithm.

## Discriminating models with measurements of representational geometry

Many studies have compared alternative models in terms of the accuracy with which they predict representational geometries, measuring how well each model predicts a given brain RDM and statistically comparing RDM prediction accuracies between models (RSA). Most of these studies have used randomly sampled or manually selected stimuli. However, a few pioneering studies have begun to explore how stimulus sets can be algorithmically selected or generated so that different models predict representational dissimilarities that are more distinct from one another<sup>30,34,38</sup>.

How can researchers best make models disagree in their predictions of representational geometry? Exact implementations of the theoretical normative objectives (Boxes 1–3) remain outstanding, but a number of effective heuristic approaches have instead been used. In each of these approaches, the stimulus set is optimized to increase some measure of distinctiveness among these predicted geometries (Fig. 1e). One heuristic is to minimize the RDM correlation for each pair of predicted RDMs<sup>30,34</sup>. For example, Groen and colleagues<sup>30</sup> used discrete optimization to minimize the correlation between the RDMs of three distinct image representations: neural network features, human-derived action labels and human-derived object labels

(Fig. 4b). They used the resulting stimulus set in behavioural and fMRI experiments, both analysed with RSA, revealing that behavioural judgements were consistent with both action labels and neural network features, whereas representations in scene-selective cortical regions were explained only by neural network features.

Jozwik and colleagues<sup>33</sup> used a genetic algorithm to disentangle five human-rated animacy dimensions ('being alive', 'looking like an animal', 'having mobility', 'having agency' and 'being unpredictable'). The five dimensions served as five models, each assigning a scalar feature value to each stimulus (Fig. 4c). A stimulus set was optimized to decorrelate the scalar model predictions and used in EEG, fMRI and behavioural experiments. RSA revealed that all dimensions explained significant perceptual variance, yet brain representations showed no trace of the latent state of 'being alive' while prominently encoding 'looking like an animal'.

Hosseini and colleagues<sup>34</sup> used stimulus selection to find a sentence set that elicits decorrelated RDMs in different language models and, similarly, an image set that elicits decorrelated RDMs in different visual neural network models. In both cases, the resulting controversial stimulus sets elicited significantly lower correlations between model RDMs and empirical RDMs, revealing shortcomings of the models that were not apparent with randomly sampled stimuli.

The RDM-decorrelation approach is intuitive, straightforward and conceptually related to maximizing the discrepancy between model predictions (Box 3). However, it neglects the effect of noise on the measured dissimilarities. Consider a scenario in which one model predicts nearly identical dissimilarities across all stimulus pairs. Such a representational geometry can arise either when the model does not distinguish any pair of stimuli (that is, all stimuli are metamers for the model) and thus the RDM contains only zeros, or when the model distinguishes all pairs of stimuli, but each pair has approximately the same dissimilarity. Even if the model is true, the correlation between its predicted RDM and the empirical RDM will be close to 0, because measurement noise will dominate the RDM correlation. We refer to this issue – where model dissimilarities are approximately equal – as the 'flat-RDM problem'. When using a correlation coefficient as the RDM comparator in model comparisons, the model RDMs need to have sufficient signal variance relative to the noise, in addition to distinct patterns of dissimilarities (here, signal variance is the variance across the predicted model dissimilarities, and noise is the error variance of the empirical dissimilarity estimates). Merely minimizing the correlation between model RDMs does not ensure that this signal-to-noise ratio is sufficient. The optimization procedure can drive two model RDMs to be uncorrelated or even anticorrelated, but if this entails a reduction of the signal variance of the dissimilarities in one or both of the model RDMs relative to the noise, then the models might not perform significantly differently at predicting the empirical RDM.

The flat-RDM problem can be avoided by optimizing a controversy objective that takes the effect of noise into account and offers a better chance of recovering the true model if it is among the models tested. If we select models on the basis of the accuracy of their RDM predictions (RSA), then the stimulus-set optimization should increase the margin by which a model that represents the ground truth achieves a higher RDM correlation with a noisy empirical RDM than other models<sup>38</sup>. For the ground-truth model to win by a significant margin, the stimulus set needs to drive distinct response patterns whose pairwise dissimilarities vary at a scale that stands out from the dissimilarity estimation errors. There are multiple sources of noise that can be considered: measurement noise, stochasticity of the predicted

responses and interparticipant variability. Interparticipant variability can be modelled as parameter uncertainty or as variability between instances of the same neural network model trained from different random seeds<sup>112</sup>.

Golan, Guo and colleagues<sup>38</sup> used controversial stimuli to compare six neural network models of face perception, taking into account variability across realizations (Fig. 3d). Each model predicted human rankings of face pairs by perceived face similarity. The models shared an architecture but differed in their training tasks. Each model was realized in two model instances by training twice from different random weight initializations. Sets of faces were synthesized by searching the space of the Basel Face Model. For a given stimulus set, the objective measured the degree to which the rankings predicted by different instances of the same model were more strongly correlated than the rankings predicted by instances of different models. Note that this optimization objective does not exactly estimate the model-recovery accuracy, but it does measure the margin by which the true model dominates competing models when measurements are affected by noise. Human judgements for the optimized set of face pairs were best predicted by a model trained to invert the generative process by inferring face latents from input images.

The approaches mentioned above compare models on the basis of their global representational geometry – the relations among the representations of clearly distinct stimuli. Alternatively, models can be compared on the basis of their local representational geometry – how the representation of a given stimulus changes when the stimulus is slightly altered<sup>40,41</sup>. Zhou and colleagues<sup>40</sup> investigated the local representational geometry by identifying a direction in input space along which a perturbation of the stimulus maximized the change of the representation in one model relative to the change of the representation in the other model (Fig. 3e). Conceptually, this approach shares the motivation of MAD competition to change the prediction of one model as much as possible while minimally affecting the prediction of the other model. Unlike MAD, this local perturbation approach relies on linearizing model responses around the response elicited by a given stimulus and can be computed for vector representations (whereas MAD is defined only for scalar outputs). In two example applications, such perturbations of local representational geometry identified shortcomings in two models of colour discrimination – motivating a third, improved model – and showed that human sensitivity to image perturbations aligned more closely with a sparse autoencoder than with a dense one. The local perturbation objective can be generalized to target more than two models<sup>41</sup> (Fig. 3f).

## Discriminating models with neural response measurements

When models predict individual neural responses (rather than representational geometries), stimuli can be optimized to make those predictions diverge. Experiments using stimuli designed to be controversial among linear encoding models have only appeared in preliminary results<sup>113,114</sup>. However, closely related approaches targeting a single model have been more deeply explored. Bashivan and colleagues<sup>66</sup> fitted linear mappings from the activations of a single neural network model to the responses of individual V4 neurons. They used this linear encoding model to generate synthetic stimuli that activate specific V4 neurons while suppressing the responses of other neurons. Given that this study aimed to control neural activity using one model rather than compare multiple models, it is not an example of model comparison with controversial stimuli. However, similar techniques could be used in the future to synthesize controversial stimuli. The controversy

objective could favour stimuli predicted by one model to activate a specific neuron but predicted by another model to suppress the same neuron (and vice versa). Alternatively, the controversiality objective could be defined to measure the extent to which a candidate stimulus drives two models to predict distinct neural activity patterns across a set of recorded neurons. Such an objective could further be generalized to multiple stimuli and multiple models, to ensure that each model can be discriminated from every other candidate model. The resulting controversial stimuli for linear encoding models may reveal differences in how well different neural network models account for a neural population code, even when the flexibility of their linear mappings allows all models to explain neural responses to natural stimuli similarly well.

A tempting choice of controversiality objective is to minimize the correlation between the response patterns predicted by alternative models. However, here as well, the controversiality objective needs to account for measurement noise. A stimulus that elicits predictions from two models that are decorrelated but not sufficiently distinct given the noise is unlikely to support accurate model selection, whether the predictions are representational dissimilarities (see the ‘Discriminating models with measurements of representational geometry’ section) or responses of individual neurons or voxels (linear encoding models).

## Outlook

We have reviewed the small number of studies that have used optimization techniques to design stimuli that make neural network models disagree in terms of either neural or behavioural predictions. These studies clearly demonstrate that controversial stimuli can reveal differences in predictive performance that remain hidden when stimuli are chosen with conventional methods such as random sampling or manual selection. However, the field is nascent. Most existing approaches aspire to optimal experimental design (Boxes 1–3) but do not attempt to implement any of the normative objectives directly. Although heuristic approaches can be effective, the field needs better methods to account for the known unknowns that determine empirical model discriminability, including measurement noise, uncertainty about model parameters, uncertainty about the model-to-brain or model-to-behaviour mapping, and variability across participants and model instances.

We close by considering how controversial stimuli can contribute to theoretical progress beyond their immediate role in increasing the power of model comparisons. We first revisit the role of out-of-distribution evaluation in testing neural networks as implementations of computational hypotheses. We then discuss two challenges that the field has only begun to address: focusing stimulus optimization on model differences of theoretical significance, and using model comparisons to guide progress despite the absence of the ‘correct’ model from the candidate set.

## Neural network models enable comparisons of computational hypotheses

Artificial neural networks can express distinct hypotheses about neural mechanisms<sup>1,5,16</sup>, yet they are also high-capacity function approximators. Their high capacity to store knowledge is essential for explaining intelligent behaviour but also makes them so flexible that it can be challenging to empirically adjudicate among them, so as to make theoretical progress. Within the training distribution, different models can converge to highly similar functions. Fortunately, despite their flexibility, two models implementing different neural mechanisms will still have distinct inductive biases that determine how they generalize

beyond the training distribution. Which is a better model of a neural mechanism can therefore be revealed by out-of-distribution challenges, where generalization transcends mere curve fitting.

A model trained to predict neural or behavioural responses to one category of natural stimuli, but evaluated on responses to another stimulus category, is making out-of-distribution predictions. Similarly, a model trained on natural stimuli but tested on artificial stimuli or stimuli optimized for model comparison is making out-of-distribution predictions. Hidden-layer representational predictions of a model trained solely on a behavioural task objective (see the ‘No flexibility: all parameters pre-fitted and considered certain’ section) also constitute an out-of-distribution generalization challenge because the model was not fitted to predict neural representations<sup>2,3</sup>. Such out-of-distribution prediction challenges can reveal the inductive biases implicit in the architecture, objective and learning rule<sup>6</sup>, even for models with millions or billions of parameters<sup>37,39</sup>.

## Finding stimuli that expose meaningful model differences

When comparing the stimulus representations learned by two or more neural network models, representational features can be conceptually grouped into two sets: features that are shared between the models and features that differ between the models. The latter set can be further subdivided into features that differ because of meaningful differences between models in terms of architecture, training task, training data or learning rule, and features that differ between models because of nuisance variability (for example, owing to different weight initializations or training-example presentation orders).

Recently, there has been a growing interest in representational similarities found in qualitatively distinct models<sup>32–34</sup>. Neural networks with different architectures or trained on different tasks not only exhibit similar predictivity of brain-activity data<sup>17,18</sup> but also acquire shared stimulus features<sup>32</sup>. It appears that features shared among models are also more likely to be shared with biological representations<sup>33</sup>. The shared features are likely to reflect regularities of the environment from which the data were sampled<sup>32</sup>. For example, object animacy often explains a considerable portion of the variance in higher-order visual representations in diverse neural network models, as well as in primate cortex (see refs. 115,116 for examples).

The observation that representational features that are shared across different models are also often shared with brains has led some authors to suggest that differences between models primarily reflect nuisance variability: two models’ non-shared features are features that the brain does not care about<sup>33,34</sup>. If this hypothesis were correct in a strong sense, with any feature not shared across models absent from brain representations, then an experiment with controversial stimuli would fail to reveal significant differences among models in their alignment with brain or behavioural data. However, when optimizing controversial stimuli for models that are qualitatively distinct – for example, models of different architectures or models trained on different tasks – the results of experiments with controversial stimuli often do reveal that certain models are markedly more aligned with brain data than others<sup>37–39</sup>. This indicates that the differences among the models do lend them distinct inductive biases and that some of the resulting distinct representational features are present in brain representations, enabling experimental adjudication among the models.

For model comparisons to drive theoretical progress, the models must implement distinct computational hypotheses. Controversial stimulus generation should capitalize on features that reflect theoretically meaningful differences between models, such as differences in

architecture, training task, training data or learning rule, rather than on features that reflect nuisance model variability, such as differences arising from random weight initializations. To ensure meaningful model controversy, it is attractive to optimize stimulus sets such that they are controversial among ensembles of models. Each ensemble represents one computational hypothesis and comprises a set of model instances that have different weights (for example, trained using different random initializations) and, optionally, variations of architecture, objective and learning rule that are not of theoretical interest. Beyond using one ensemble of models per computational hypothesis, researchers can also evaluate continuous spaces of computational hypotheses in this framework. The general requirement is that the set of models should capture both variation of theoretical interest (such as qualitatively distinct architectures) and nuisance variation (such as small changes in architectural hyperparameters), and analyses should reveal how the factors of theoretical interest affect predictive performance.

In the longer term, it will be essential that researchers develop distributions of models that capture remaining uncertainty about the computational mechanisms in a more comprehensive fashion. This includes representing model parameters as random variables<sup>117–119</sup>, as well as capturing not only group-mean tendencies but also inter-individual variability in neural representations and behaviour. Ideally, neuroscientists should compare theories corresponding to distributions of mechanistic models with data sampled from distributions of individual animals. Controversial stimuli should therefore generalize across the idiosyncrasies of model instances and individuals. The model-comparative power of such stimulus sets will be measured by their ability to discriminate among ensembles of models, even when new model instances are used to generate the predictions and responses are sampled from new individual participants.

## Theoretical progress when the true model is elusive

Each candidate model represents a hypothesis about the mechanism underlying a particular cognitive function. Ideally, researchers would discover the model (or model ensemble) that implements the true mechanism generating the neural and behavioural responses to the stimuli. This true model would make accurate predictions of brain representations across stages of processing and behavioural responses for all stimuli, natural or artificial. However, for a given study that tries to make progress on a complex cognitive function such as visual recognition or spatial navigation, it is unrealistic to assume that the true model is among the models under consideration. Some models may be closer to the truth than others, but in general, the true model (the actual data-generating mechanism) will not be among the candidate models tested.

All models are wrong<sup>120</sup>, especially in cognitive computational neuroscience. This reflects both the stage of development of the field and the intrinsic complexity of brains. Cognitive computational neuroscience finds itself in what is known as the open-set, or  $\mathcal{M}$ -open scenario (see section 6.1.2 in ref. 121), which is distinct from the closed-set system-identification framework assumed in optimal experimental design (Boxes 1–3). The absence of the true model precludes strict Bayesian inference. If all models are misspecified, researchers also cannot get to the truth by refuting alternatives. Instead, the field must navigate the space of models seeking progress towards models that are less incorrect<sup>8,16,122,123</sup>. This motivates a focus on statistical comparison of alternative imperfect models.

A consequent challenge is that different models may dominate for different sets of stimuli: two experiments with different stimulus sets

may yield apparently inconsistent model comparison results. Model A may perform significantly better than model B in experiment 1, and model B may perform significantly better than model A in experiment 2; thus neither model dominates the other. If the true model were among the models, it would not be beaten by any other model, no matter what stimulus set was used for the comparison. In particular, the true model would be among the winning models for naturalistic as well as non-naturalistic stimuli. However, in the absence of the true model, there is only a set of imperfect models, each exhibiting a specific pattern of systematic prediction errors. A particular experiment will reveal model performance for a limited portion of stimulus space in which one model may do better than others even if it fails badly elsewhere. For example, Golan and colleagues<sup>37</sup> found that discriminative neural networks predicted human recognition of handwritten digits slightly more accurately than alternative models when tested with standard MNIST digits. However, when synthetic controversial stimuli were used, a generative neural network<sup>124</sup> predicted human digit recognition substantially more accurately (Fig. S8 in ref. 37).

To eliminate ambiguities in model ranking, researchers might be tempted to commit to a single stimulus set, such as a large set of natural stimuli or a combination of all natural and artificial stimulus sets for which neural or behavioural response measurements are available. Using a single stimulus set, researchers can combine all the evidence and reduce each model's performance profile to a single number. This approach will not falsely eliminate the true model. However, it tacitly commits to a relative weighting of different portions of stimulus space and may misguide researchers in assessing the relative merits of imperfect models. Models that do not perform well overall may excel on particular stimuli and may even displace the current winners when a different weighting of the datasets is chosen or additional datasets are added to the benchmark, such that the distribution of test stimuli changes.

Rather than prematurely reducing complex profiles of performance to scalar scores, researchers can embrace multiple evaluations of each model across stimulus sets and prediction targets (such as brain regions and behavioural tasks). Embracing multiple evaluations enables researchers to gauge the scope of stimuli and the scope of responses over which each model predicts well. This information is essential in choosing models for particular applications. Analyses of where and how models fail can also provide guidance for improving models (see ref. 125, for example). Explainable artificial intelligence methods can help researchers to understand why failures occur and how a model might be improved. Approaching computational modelling as a multi-objective optimization problem, rather than a contest on a single benchmark, ensures that distinct failure modes are not masked by aggregation.

An important tool for handling the ambiguity that comes with multiple objectives is the concept of Pareto optimality. Rather than attempting to identify a single winning model, a Pareto set can be identified. A model is in the Pareto set if it is non-dominated – that is, if no single other model performs at least as well across all objectives (all stimulus sets and/or prediction targets) and better on at least one objective. Modelling should aim to narrow the Pareto set of non-dominated models, which requires some models to dominate others. However, new experiments may uncover new model failure modes, which can bring non-competitive (dominated) models back into the Pareto set. Identifying a Pareto set instead of aggregating model performance avoids the arbitrary weighting of different stimulus sets and prediction targets.

## Glossary

### Adversarial examples

An input pattern deliberately constructed to elicit an incorrect response from a model; for example, a natural image subtly modified by a human-imperceptible perturbation discovered through gradient descent to cause the model to assign the highest probability to an incorrect class label.

### Controversial stimuli

Input patterns deliberately chosen or generated to elicit divergent neural or behavioural response predictions from competing models, so as to enable efficient empirical comparison among the models.

### Convolutional neural network

A neural network architecture that learns identical filters with shared parameters across locations (for example, across space in images or across time in audio), yielding representations that translate as the input is translated and automatically generalize across locations.

### Critical experiments

Experiments designed to adjudicate among competing theories by testing cases where those theories make clearly different predictions.

### $k$ -alternative forced choice

( $k$ -AFC). A behavioural task in which a participant selects one of  $k$  alternative stimuli on each trial (for example, an image, a location, or a temporal interval that stands out along a specified perceptual dimension), such that the participant does not need to impose a subjective threshold for the decision as is required in choosing one of  $k$  abstract response categories such as 'yes/no' or 'same/different' ( $k = 2$ ).

### Kullback–Leibler (KL) divergence

An asymmetric measure  $D_{\text{KL}}(P\|Q)$  of how much probability distribution  $P$  differs from distribution  $Q$ , which measures the expected additional surprise experienced by a believer in  $Q$  when witnessing events drawn from  $P$ , compared to a believer in the actual distribution  $P$ .

### Latent factors

Unobserved variables inferred from data to explain the statistical structure of the observed variables.

### Manifold

A possibly curved, low-dimensional subspace (for example, all face images) of a larger space (all pixel images), such that the subspace is locally homeomorphic to a Euclidean space at every point.

### Model capacity

The representational flexibility afforded by a model's adjustable parameters, which generally grows with the number of independently adjustable parameters but is reined in by any prior over the parameters (or regularization) and can be formalized by function-class complexity or information-theoretic measures.

### Model-recovery accuracy

The accuracy with which a given method for model adjudication recovers the true model, estimated as the proportion of correct model-selection decisions made on the basis of simulated data where the true model is known.

### Neural network-based generator

A neural network model trained to generate new samples (for example, visual images) from a complex distribution (for example, the distribution of human faces); for example, by using a diffusion or generative adversarial-training approach.

### Neural network models

Computational models composed of units inspired by biological neurons but highly abstracted, where each unit typically computes a linear combination of its inputs with learned weights, followed by a static nonlinearity, and units are connected to form a network of feedforward or recurrent transformations.

### Regularization

Any method that penalizes certain settings of the adjustable parameters of a model (such as large weights) during fitting so as to reduce overfitting and improve generalization to new data.

### Transformer neural network

A neural network architecture that uses a dynamic connectivity mechanism called attention that lets elements of a set of inputs interact with one another so as to flexibly contextualize each element with information from other related elements.

### Universal function approximators

A model class that can approximate any function in a specified broad class arbitrarily closely, given sufficient parameters and appropriate settings of those parameters.

### Universal representation hypothesis

The idea that high-capacity models trained on rich tasks may converge on similar internal representations that reflect the latent structure of the world, even if their architectures or training regimes differ substantially.

Although the field should strive for unified mechanistic explanations that make correct predictions under a wide range of conditions (including natural stimuli of different categories, as well as adversarial and controversial stimuli), particular studies must compare imperfect models that are typically far from offering unified explanations. To navigate this gap between the ideal and the real, researchers can adjust the scope of their explanatory ambition, narrowing or expanding the stimulus domain and range of prediction targets across which candidate models are expected to perform well. By defining the space of admissible stimuli for experimental design, they can choose a scope that pushes towards more unified explanations while acknowledging the limitations of current models.

## Conclusion

Cognitive computational neuroscience is in an exciting phase of growth and insight but still lacks a mature empirical methodology. Techniques

for designing experiments to adjudicate among computational hypotheses will be a key component of this methodology. Despite many challenges, recent advances with controversial stimuli give researchers much power to compare and reject neural network models, enabling them to navigate the vast universe of computational hypotheses.

Published online: 28 August 2026

## References

1. O'Reilly, R. C. & Munakata, Y. *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain* (MIT Press, 2000).
2. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
3. Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).
4. Wu, M. C.-K., David, S. V. & Gallant, J. L. Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.* **29**, 477–505 (2006).
5. Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nat. Neurosci.* **21**, 1148–1160 (2018).

6. Richards, B. A. et al. A deep learning framework for neuroscience. *Nat. Neurosci.* **22**, 1761–1770 (2019).
7. Saxe, A., Nelli, S. & Summerfield, C. If deep learning is the answer, what is the question? *Nat. Rev. Neurosci.* **22**, 55–67 (2021).
8. Doerig, A. et al. The neuroconnectionist research programme. *Nat. Rev. Neurosci.* **24**, 431–450 (2023).
9. DePasquale, B., Sussillo, D., Abbott, L. F. & Churchland, M. M. The centrality of population-level factors to network computation is demonstrated by a versatile approach for training spiking networks. *Neuron* **111**, 631–649.e10 (2023).
10. Langdon, C., Genkin, M. & Engel, T. A. A unifying perspective on neural manifolds and circuits for cognition. *Nat. Rev. Neurosci.* **24**, 363–377 (2023).
11. Felsen, G. & Dan, Y. A natural approach to studying vision. *Nat. Neurosci.* **8**, 1643–1646 (2005).
12. Vinje, W. E. & Gallant, J. L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).
13. Vinje, W. E. & Gallant, J. L. Natural stimulation of the nonclassical receptive field increases information transmission efficiency in V1. *J. Neurosci.* **22**, 2904–2915 (2002).
14. Prenger, R., Wu, M. C.-K., David, S. V. & Gallant, J. L. Nonlinear V1 responses to natural scenes revealed by neural network analysis. *Neural Netw.* **17**, 663–679 (2004).
15. Rust, N. C. & Movshon, J. A. In praise of artifice. *Nat. Neurosci.* **8**, 1647–1650 (2005).
16. Golan, T. et al. Deep neural networks are not a single hypothesis but a language for expressing computational hypotheses. *Behav. Brain Sci.* **46**, e392 (2023).
17. Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J. & Kriegeskorte, N. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *J. Cogn. Neurosci.* **33**, 2044–2064 (2021).
18. Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A. & Konkle, T. A large-scale examination of inductive biases shaping high-level visual representation in brains and machines. *Nat. Commun.* **15**, 9383 (2024).
19. Schrimpf, M. et al. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* **108**, 413–423 (2020).
20. Brain-Score <https://brain-score.org> (accessed 2 February 2026).
21. Han, Y., Poggio, T. A. & Cheung, B. System identification of neural systems: if we got it right, would we know? In *International Conference on Machine Learning*, Vol. 202 of *Proc. Machine Learning Research* <https://proceedings.mlr.press/v202/han23d.html> (PMLR, 2023).
22. Kubilius, J. et al. Brain-like object recognition with high-performing shallow recurrent ANNs. In *Advances in Neural Information Processing Systems* [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/7813d1590d28a7dd372ad54b5d29d033-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/7813d1590d28a7dd372ad54b5d29d033-Paper.pdf) (Curran Associates, 2019).
23. Dapello, J. et al. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. In *Advances in Neural Information Processing Systems* [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/98b17f068d5d9b7668e19fb8ae470841-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/98b17f068d5d9b7668e19fb8ae470841-Paper.pdf) (Curran Associates, 2020).
24. Kietzmann, T. C. et al. Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl Acad. Sci. USA* **116**, 21854–21863 (2019).
25. Chang, L., Egger, B., Vetter, T. & Tsao, D. Y. Explaining face representation in the primate brain using different computational models. *Curr. Biol.* **31**, 2785–2795.e4 (2021).
26. Grossman, S. et al. Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* **10**, 4934 (2019).
27. Carlin, J. D. & Kriegeskorte, N. Adjudicating between face-coding models with individual-face fMRI responses. *PLoS Comput. Biol.* **13**, e1005604 (2017).
28. Jozwik, K. M. et al. Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. *Proc. Natl Acad. Sci. USA* **119**, e2115047119 (2022).
29. Wang, Z. & Simoncelli, E. P. Maximum differentiation (MAD) competition: a methodology for comparing computational models of perceptual quantities. *J. Vis.* **8**, 8–8 (2008).
30. Groen, I. I. et al. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife* **7**, e32962 (2018).
31. Jozwik, K. M. et al. Disentangling five dimensions of animacy in human brain and behaviour. *Commun. Biol.* **5**, 1247 (2022).
32. Huh, M., Cheung, B., Wang, T. & Isola, P. Position: the platonic representation hypothesis. In *International Conference on Machine Learning*, Vol. 235 of *Proceedings of Machine Learning Research* <https://proceedings.mlr.press/v235/huh24a.html> (PMLR, 2024).
33. Chen, Z. & Bonner, M. F. Universal dimensions of visual representation. *Sci. Adv.* **11**, eadw7697 (2025).
34. Hosseini, E. et al. Universality of representation in biological and artificial neural networks. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.12.26.629294> (2024).
35. Avitan, I. & Golan, T. Model–behavior alignment under flexible evaluation: when the best-fitting model isn't the right one. In *Advances in Neural Information Processing Systems* [https://proceedings.neurips.cc/paper\\_files/paper/2025/file/11e1900e680f5fe1893a8e27362dbe2c-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2025/file/11e1900e680f5fe1893a8e27362dbe2c-Paper-Conference.pdf) (Curran Associates, 2025).
36. Javed, K. & Sutton, R. S. The big world hypothesis and its ramifications for artificial intelligence. In *Finding the Frame: an RLC Workshop for Examining Conceptual Frameworks* <https://openreview.net/forum?id=sv7DazuCn8> (OpenReview.net, 2024).
37. Golan, T., Raju, P. C. & Kriegeskorte, N. Controversial stimuli: pitting neural networks against each other as models of human cognition. *Proc. Natl Acad. Sci. USA* **117**, 29330–29337 (2020).
38. Golan, T., Guo, W., Schütt, H. H. & Kriegeskorte, N. Distinguishing representational geometries with controversial stimuli: Bayesian experimental design and its application to face dissimilarity judgments. In *SVRHM 2022 Workshop @ NeurIPS* <https://openreview.net/forum?id=a3Ypu2-Mf2h> (OpenReview.net, 2022).
39. Golan, T., Siegelman, M., Kriegeskorte, N. & Baldassano, C. Testing the limits of natural language models for predicting human language judgements. *Nat. Mach. Intell.* **5**, 952–964 (2023).
40. Zhou, J., Chun, C., Subramanian, A. & Simoncelli, E. P. Comparing neural models using their perceptual discriminability predictions. In *Proc. UniReps: the First Workshop on Unifying Representations in Neural Models*, Vol. 243 of *Proc. Machine Learning Research* <https://proceedings.mlr.press/v243/zhoul24a.html> (PMLR, 2024).
41. Feather, J., Lipshutz, D., Harvey, S. E., Williams, A. H. & Simoncelli, E. P. Discriminating image representations with principal distortions. In *International Conference on Learning Representations* <https://openreview.net/forum?id=ugXGFCs6HK> (OpenReview.net, 2025).
42. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).
43. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
44. Nili, H. et al. A toolbox for representational similarity analysis. *PLoS Comput. Biol.* **10**, e1003553 (2014).
45. Diedrichsen, J. & Kriegeskorte, N. Representational models: a common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput. Biol.* **13**, e1005508 (2017).
46. Diedrichsen, J. et al. Comparing representational geometries using whitened unbiased-distance-matrix similarity. *Neurons Behav. Data Anal. Theory* <https://doi.org/10.51628/001c.27664> (2021).
47. Schütt, H. H., Kipnis, A. D., Diedrichsen, J. & Kriegeskorte, N. Statistical inference on representational geometries. *eLife* **12**, e25666 (2023).
48. Williams, A. H., Kunz, E., Kornblith, S. & Linderman, S. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems* <https://proceedings.neurips.cc/paper/2021/hash/252a3dbae332e7690242ad3b556e626b-Abstract.html> (Curran Associates, 2021).
49. Williams, A. H. Equivalence between representational similarity analysis, centered kernel alignment, and canonical correlations analysis. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models* <https://openreview.net/forum?id=zMdnFasgC> (OpenReview.net, 2024).
50. van den Bosch, J. J. F. et al. A Python toolbox for representational similarity analysis. Preprint at *eLife* <https://doi.org/10.7554/eLife.107828.1> (2025).
51. St-Yves, G. & Naselaris, T. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *Neuroimage* **180**, 188–202 (2018).
52. Kriegeskorte, N. & Diedrichsen, J. Peeling the onion of brain representations. *Annu. Rev. Neurosci.* **42**, 407–432 (2019).
53. Kriegeskorte, N. & Wei, X.-X. Neural tuning and representational geometry. *Nat. Rev. Neurosci.* **22**, 703–718 (2021).
54. Paninski, L. Maximum likelihood estimation of cascade point-process neural encoding models. *Network* **15**, 243–262 (2004).
55. Peterson, J. C., Abbott, J. T. & Griffiths, T. L. Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cogn. Sci.* **42**, 2648–2669 (2018).
56. Battleday, R. M., Peterson, J. C. & Griffiths, T. L. Capturing human categorization of natural images by combining deep networks and cognitive models. *Nat. Commun.* **11**, 5418 (2020).
57. Muttenthaler, L., Dippel, J., Linhardt, L., Vandermeulen, R. A. & Kornblith, S. Human alignment of neural network representations. In *International Conference on Learning Representations* <https://openreview.net/forum?id=ReDQ1OUQROX> (OpenReview.net, 2023).
58. Kriegeskorte, N. & Diedrichsen, J. Inferring brain-computational mechanisms with models of activity measurements. *Philos. Trans. R. Soc. B* **371**, 20160278 (2016).
59. Friston, K. J., Diedrichsen, J., Holmes, E. & Zeidman, P. Variational representational similarity analysis. *Neuroimage* **201**, 115986 (2019).
60. Diedrichsen, J., Ridgway, G. R., Friston, K. J. & Wiestler, T. Comparing the similarity and spatial structure of neural representations: a pattern-component model. *Neuroimage* **55**, 1665–1678 (2011).
61. Lin, B. & Kriegeskorte, N. The topology and geometry of neural representations. *Proc. Natl Acad. Sci. USA* **121**, e2317881121 (2024).
62. Kornblith, S., Norouzi, M., Lee, H. & Hinton, G. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, Vol. 97 of *Proc. Machine Learning Research* <https://proceedings.mlr.press/v97/kornblith19a.html> (PMLR, 2019).
63. Walther, A. et al. Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage* **137**, 188–200 (2016).
64. Murphy, A. G., Zylberberg, J. & Fyshe, A. Correcting biased centered kernel alignment measures in biological and artificial neural networks. In *ICLR 2024 Workshop on Representational Alignment* <https://openreview.net/forum?id=E1NRgtlHG> (OpenReview.net, 2024).
65. Ivanova, A. A. et al. Beyond linear regression: mapping models in cognitive neuroscience should align with research goals. *Neurons Behav. Data Anal. Theory* <https://doi.org/10.51628/001c.37507> (2022).

66. Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
67. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
68. Khaligh-Razavi, S.-M., Henriksson, L., Kay, K. & Kriegeskorte, N. Fixed versus mixed RSA: explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *J. Math. Psychol.* **76**, 184–197 (2017).
69. Jozwik, K. M., Kriegeskorte, N. & Mur, M. Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia* **83**, 201–226 (2016).
70. Kaniuth, P. & Hebart, M. N. Feature-reweighted representational similarity analysis: a method for improving the fit between computational models, brains, and behavior. *Neuroimage* **257**, 119294 (2022).
71. Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
72. Schütt, H. H. Bayesian comparisons between representations. In *Proc. Cognitive Computational Neuroscience 2025*; <https://doi.org/10.32470/050h4pji> (Cognitive Computational Neuroscience, 2025).
73. Walker, E. Y. et al. Inception loops discover what excites neurons most using deep predictive models. *Nat. Neurosci.* **22**, 2060–2065 (2019).
74. Seeliger, K. et al. End-to-end neural system identification with neural information flow. *PLoS Comput. Biol.* **17**, e1008558 (2021).
75. Khosla, M., Ngo, G. H., Jamison, K., Kuceyeski, A. & Sabuncu, M. R. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Sci. Adv.* **7**, eabe7547 (2021).
76. St-Yves, G., Allen, E. J., Wu, Y., Kay, K. & Naselaris, T. Brain-optimized deep neural network models of human visual areas learn non-hierarchical representations. *Nat. Commun.* **14**, 3329 (2023).
77. Lindley, D. V. On a measure of the information provided by an experiment. *Ann. Math. Stat.* **27**, 986–1005 (1956).
78. Chaloner, K. & Verdinelli, I. Bayesian experimental design: a review. *Stat. Sci.* **10**, 273–304 (1995).
79. Paninski, L. Asymptotic theory of information-theoretic experimental design. *Neural Comput.* **17**, 1480–1507 (2005).
80. Box, G. E. P. & Hill, W. J. Discrimination among mechanistic models. *Technometrics* **9**, 57–71 (1967).
81. Atkinson, A. C. & Fedorov, V. V. The design of experiments for discriminating between two rival models. *Biometrika* **62**, 57–70 (1975).
82. Atkinson, A. C. & Fedorov, V. V. Optimal design: experiments for discriminating between several models. *Biometrika* **62**, 289–303 (1975).
83. Dette, H. & Tioff, S. Optimal discrimination designs. *Ann. Stat.* **37**, 2056–2082 (2009).
84. Dette, H., Melas, V. B. & Shpilev, P. T-optimal designs for discrimination between two polynomial models. *Ann. Stat.* **40**, 188–205 (2012).
85. Guo, C. et al. Adversarially trained neural representations may already be as robust as corresponding biological neural representations. In *International Conference on Machine Learning* <https://proceedings.mlr.press/v162/guo22d.html> (PMLR, 2022).
86. Greene, M. R. & Hansen, B. C. Function over form: the temporal evolution of affordance-based scene categorization. *J. Vis.* **25**, 10–10 (2025).
87. Ponce, C. R. et al. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* **177**, 999–1009.e10 (2019).
88. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, 2006).
89. Blanz, V. & Vetter, T. A morphable model for the synthesis of 3D faces. In *Proc. 26th annual conference on Computer graphics and interactive techniques — SIGGRAPH '99* <http://portal.acm.org/citation.cfm?doid=311535.311556> (ACM Press, 1999).
90. Daube, C. et al. Grounding deep neural network predictions of human categorization behavior in understandable functional features: the case of face identity. *Patterns* **2**, 100348 (2021).
91. Karras, T. et al. Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems* [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/076ccd93ad68be51f23707988e934906-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/076ccd93ad68be51f23707988e934906-Paper.pdf) (Curran Associates, 2021).
92. Papantoniou, F. P. et al. Arc2Face: a foundation model for ID-consistent human faces. In *Computer Vision — ECCV 2024*, 241–261 (Springer, 2025).
93. Luo, A., Henderson, M., Wehbe, L. & Tarr, M. Brain diffusion for visual exploration: Cortical discovery using large scale generative models. In *Advances in Neural Information Processing Systems* (Curran Associates, 2023); [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/ef0c0a23a1a8219c4fc381614664df3e-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/ef0c0a23a1a8219c4fc381614664df3e-Paper-Conference.pdf).
94. Olah, C., Mordvintsev, A. & Schubert, L. Feature visualization. *Distill* <https://doi.org/10.23915/distill.00007> (2017).
95. Andradóttir, S. A global search method for discrete stochastic optimization. *SIAM J. Optim.* **6**, 513–530 (1996).
96. Doerr, B. & Neumann, F. A survey on recent progress in the theory of evolutionary algorithms for discrete optimization. *ACM Trans. Evol. Learn. Optim.* **1**, 1–43 (2021).
97. Rainforth, T., Foster, A., Ivanova, D. R. & Bickford Smith, F. Modern Bayesian experimental design. *Stat. Sci.* <https://doi.org/10.1214/23-ST915> (2024).
98. Goda, T., Hironaka, T., Kitade, W. & Foster, A. Unbiased MLMC stochastic gradient-based optimization of Bayesian experimental designs. *SIAM J. Sci. Comput.* **44**, A286–A311 (2022).
99. Foster, A. et al. Variational Bayesian optimal experimental design. In *Advances in Neural Information Processing Systems* [https://proceedings.neurips.cc/paper\\_files/paper/2019/hash/d55cbf210f175f4a37916eafe6c04f0d-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2019/hash/d55cbf210f175f4a37916eafe6c04f0d-Abstract.html) (Curran Associates, 2019).
100. Foster, A., Ivanova, D. R., Malik, I. & Rainforth, T. Deep adaptive design: amortizing sequential Bayesian experimental design. In *International Conference on Machine Learning* <https://proceedings.mlr.press/v139/foster21a.html> (PMLR, 2021).
101. Myung, J. I. & Pitt, M. A. Optimal experimental design for model discrimination. *Psychol. Rev.* **116**, 499–518 (2009).
102. Tuckute, G. et al. Driving and suppressing the human language network using large language models. *Nat. Hum. Behav.* **8**, 544–561 (2024).
103. Berardino, A., Laparra, V., Ballé, J. & Simoncelli, E. Eigen-distortions of hierarchical representations. In *Advances in Neural Information Processing Systems* [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/c5a4e7e6882845ea7bb4d9462868219b-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/c5a4e7e6882845ea7bb4d9462868219b-Paper.pdf) (Curran Associates, 2017).
104. Veerabadran, V. et al. Subtle adversarial image manipulations influence both human and machine perception. *Nat. Commun.* **14**, 4933 (2023).
105. Yuan, L. et al. Fooling the primate brain with minimal, targeted image manipulation. Preprint at <https://doi.org/10.48550/arXiv.2011.05623> (2020).
106. Freeman, J. & Simoncelli, E. P. Metamers of the ventral stream. *Nat. Neurosci.* **14**, 1195–1201 (2011).
107. Feather, J., Leclerc, G., Madry, A. & McDermott, J. H. Model metamers reveal divergent invariances between biological and artificial neural networks. *Nat. Neurosci.* **26**, 2017–2034 (2023).
108. Storrs, K. R., Anderson, B. L. & Fleming, R. W. Unsupervised learning predicts human perception and misperception of gloss. *Nat. Hum. Behav.* **5**, 1402–1417 (2021).
109. Gomez-Villa, A., Martin, A., Vazquez-Corral, J., Bertalmio, M. & Malo, J. On the synthesis of visual illusions using deep generative models. *J. Vis.* **22**, 2 (2022).
110. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
111. Ma, K. et al. Group MAD competition — a new methodology to compare objective image quality models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 1664–1673 (IEEE, 2016).
112. Mehrer, J., Spoerer, C. J., Kriegeskorte, N. & Kietzmann, T. C. Individual differences among deep neural network models. *Nat. Commun.* **11**, 5725 (2020).
113. Skril, D., Feather, J. & Norman-Haignere, S. V. Universally controversial stimuli reveal that adversarial robustness improves DNN prediction accuracy across the entire human auditory cortex. In *Conference on Cognitive Computational Neuroscience* <https://2025.ccneuro.org/poster/?id=VQ31hJQ4O4> (CCN, 2025).
114. Prince, J. S. et al. Parametric control along the encoding axes of IT neurons uncovers hidden differences in model-brain alignment. In *Conference on Cognitive Computational Neuroscience* <https://2025.ccneuro.org/poster/?id=XHkMEpOnhl> (CCN, 2025).
115. Kriegeskorte, N. et al. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–1141 (2008).
116. Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
117. Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. Weight uncertainty in neural networks. In *International Conference on Machine Learning* <https://proceedings.mlr.press/v37/blundell15.html> (PMLR, 2015).
118. McClure, P. & Kriegeskorte, N. Robustly representing uncertainty through sampling in deep neural networks. In *Bayesian Deep Learning, NIPS 2017 Workshop* <https://bayesandeeplearning.org/2017/papers/21.pdf> (OpenReview.net, 2017).
119. Gal, Y. & Ghahramani, Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In *International Conference on Machine Learning* <https://proceedings.mlr.press/v48/gal16.html> (PMLR, 2016).
120. Box, G. E. P. Science and statistics. *J. Am. Stat. Assoc.* **71**, 791–799 (1976).
121. Bernardo, J. M. & Smith, A. F. M. *Bayesian Theory* (Wiley, 2004).
122. Popper, K. R. *The Logic of Scientific Discovery* 2nd edn (Routledge, 2005).
123. Lakatos, I. *The Methodology of Scientific Research Programmes: Philosophical Papers* <https://doi.org/10.1017/CBO9780511621123> (Cambridge Univ. Press, 1978).
124. Schott, L., Rauber, J., Bethge, M. & Brendel, W. Towards the first adversarially robust neural network model on MNIST. In *International Conference on Learning Representations* <https://openreview.net/forum?id=S1EHOsc9TX> (OpenReview.net, 2019).
125. D'Agostino, F., Schwetlick, L., Bethge, M. & Kuemmerer, M. What moves the eyes: doubling mechanistic model performance using deep networks to discover and test cognitive hypotheses. In *Advances in Neural Information Processing Systems* <https://openreview.net/forum?id=cnrwlqr8dg> (Curran Associates, 2025).
126. Hainy, M., Price, D. J., Restif, O. & Drovandi, C. Optimal Bayesian design for model discrimination via classification. *Stat. Comput.* **32**, 25 (2022).
127. López-Fidalgo, J., Tommasi, C. & Trandafir, P. C. An optimal experimental design criterion for discriminating between non-normal models. *J. R. Stat. Soc. B Stat. Methodol.* **69**, 231–242 (2007).

## Acknowledgements

The authors thank L. Paninski for his valuable feedback on an early draft of this manuscript. This work was supported by the National Institute of Neurological Disorders and Stroke of the National Institutes of Health under Award Number RFINS128897 to N.K., and by the Israel Science Foundation (grant no. 534/24) to T.G. The content is solely the responsibility of the

---

# Review article

---

authors and does not necessarily represent the official views of the National Institutes of Health or the Israel Science Foundation.

## Author contributions

T.G. and N.K. researched data for and wrote the article. All authors contributed substantially to discussion of the content and edited the manuscript before submission.

## Competing interests

The authors declare no competing interests.

## Additional information

**Peer review information** *Nature Reviews Neuroscience* thanks Brian Cheung and the other, anonymous, reviewer(s) 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.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature Limited 2026