Brain-like variability in convolutional neural networks reveals evidence-, uncertainty- and bias-driven decision-making

Johan Nakuci[1*]

[1]U.S. Army DEVCOM Army Research Laboratory, Aberdeen Proving Ground, Aberdeen, MD, 21005, USA

Email: jnakuci@gmail.com

Competing Interest Statement: Authors declare that they have no competing interests.

Keywords: Convolutional neural networks, trial-to-trial variability, perceptual decision-making, default mode network, bias

Author Contributions:
Conceptualization: JN
Methodology: JN
Data Curation: JN
Visualization: JN
Funding acquisition: JN
Writing – original draft: JN
Writing – review & editing: JN

Abstract

Brain activity fluctuates from trial to trial even when task demands are constant, yet how such variability shape decisions remains unresolved. Using a convolutional neural network trained on a perceptual decision task, we show that reliable changes in decision-making processes. By decomposing the decision variable into stimulus-dependent activity and stimulus-independent bias, we identify states in which bias is disproportionately influential—especially when evidence is weak or ambiguous. We further introduce a layer-resolved Bias Dominance Index (BDI) that quantifies this balance across network depth and generalizes across architectures and datasets. These findings suggest that when stimulus-driven signals are attenuated decisions may remain stable by shifting weight toward stimulus-independent bias. Together, these findings show that variability in CNN activity can serve as a model for determining how the brain transforms changes in internal states into adaptive behavior, providing a bridge between fluctuations in activity, decision-making, and behavior.

Introduction

Brain activity exhibits pervasive trial-to-trial variability despite consistent task demands (1, 2). Although this variability is often treated as noise in decision-making models, converging evidence suggests that fluctuations can support adaptive behavior in changing environments (3). A central unresolved question is how moment-to-moment changes in brain activity translate into shifts in decisions, particularly in cases where markedly different internal activity patterns yield similar observable behavior.

This gap is underscored by findings from perceptual decision-making tasks in which brain regions can exhibit stark variation in activity with minimal impact on performance (4). In our prior work, a subset of trials showed increased engagement of the Default Mode Network — brain regions commonly associated with mind-wandering (5) and being off task (6) — yet accuracy, reaction times, and confidence were comparable to trials in which the Default Mode Network was less engaged and regions linked to focused attention were prominent (**Fig. 1A**). This dissociation motivates the question of how can decisions remain stable when stimulus-driven engagement appears reduced, and what computations compensate when sensory evidence is weak or ambiguous.

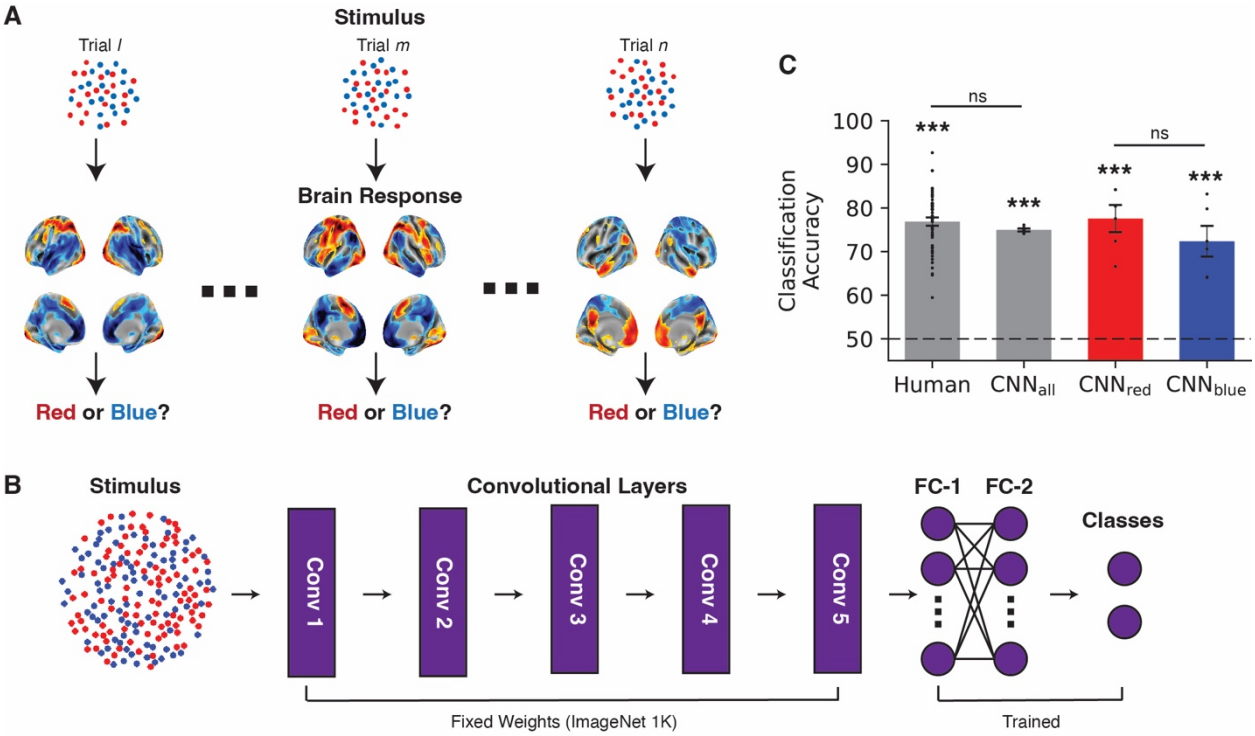

Figure 1. Multiple patterns of activation and CNN implementation of a perceptual decision-making task. (A) Canonical multiple activation patterns identified during a perceptual decision-making task. Subjects were tasked with identifying the dominant color in an image consisting of red and blue dots. Brain maps were originally published in Nakuci et al., 2025(4). (B) Schematic of CNN architecture. During training only the weights of the fully connected layers were modified. Whereas the weights in the convolution component were fixed based on ImageNet. (C) CNN and Human classification performance. Human data is based on 50 subjects from Nakuci et al., 2025(4). Bar plots depict the mean ± sem. Black dots depict the performance of individuals and models, respectively. The dashed line indicates chance performance (50%). ***P < 0.001, ns = not significant.

Disentangling the origins and functional consequences of trial-to-trial variability in the brain is difficult because fluctuations can reflect multiple interacting sources, including top-down processes related to attention (7), recurrent dynamics within local circuits (8), and bottom-up stimulus-driven computations (9) all measured through inherently noisy observation processes. This complexity motivates complementary approaches in systems where the full internal state is accessible, and the sources of variability can be controlled. Convolutional neural networks (CNNs) provide such a setting. When trained on objectives that approximate sensory categorization demands, CNNs have proven useful for understanding object recognition (10, 11), representational structure (12) and aspects of processing in the visual cortex (13–15). Here, we leverage this tractability to probe how decision computations change as internal activations fluctuate across trials.

Specifically, we train CNNs to perform a perceptual decision-making task—identifying the dominant color in images containing red and blue dots—and examine trial-level activation patterns throughout the network. Because the model is deterministic and trial-to-trial differences can be driven by controlled stimulus variation, this approach isolates a baseline contribution of stimulus-driven dynamics to variability in internal activity. We then introduce a mechanistic decomposition of the decision variable into stimulus-dependent evidence and stimulus-independent bias, allowing us to quantify when decisions are primarily evidence-driven versus bias-weighted. To test whether bias contributions are confined to the final readout or emerge across processing stages, we develop a layer-resolved Bias Dominance Index (BDI) that tracks the balance between evidence and bias across network depth and generalizes across architectures and datasets. Together, this framework provides a tractable computational account of how distinct decision modes can arise from the same task objective, and it yields testable predictions for behavioral modes in which reduced stimulus-driven engagement—potentially including DMN-

associated periods—may be accompanied by increased reliance on bias-like components of the decision computation.

Results

Implementation of a perceptual decision-making task in a CNN

To test how trial-level fluctuations in internal activity relate to decision-making, we trained CNNs on a perceptual decision task previously used to study trial-level variability in humans (4). Each CNN reported the dominant color (red vs blue) in images containing intermixed red and blue dots (**Fig. 1B**). We trained and evaluated five independently trained models.

Across models, CNNs performed significantly above chance (Accuracy = 74.96 ± 0.38%; $t(4) = 65.41$, $P = 3.27 \times 10^{-}$) and did not differ significantly from human accuracy on the same task (Accuracy = 76.87 ± 0.95%; $t = 1.84$, $P = 0.07$; **Fig. 1C**). Performance was also above chance when evaluated separately for red-dominant trials (Accuracy = 77.57 ± 3.48%; $t(4) = 7.91$, $P = 1.38 \times 10^{-3}$) and blue-dominant trials (Accuracy = 72.36 ± 3.95%; $t(4) = 5.65$, $P = 0.004$), with no evidence for a systematic preference for either dominant color ($t(4) = 0.70$, $P = 0.52$). These results establish that the models learned the task and support trial-level analyses of internal representations.

CNN contains multiple but stable trial modes during perceptual decision-making

We focused subsequent analyses on the penultimate layer (FC-2; **Fig. 1B**), which provides the internal representation used by the classifier and therefore directly shapes the decision output. To determine whether trial-level activity organizes into reproducible patterns, we applied modularity-maximization clustering to trial-by-trial similarity matrices computed from FC-2 activations, using Pearson correlation over neurons active on at least one trial (16).

In a representative model (Model-4), clustering identified three subtypes (**Fig. 2A**). Subtype 1 accounted for 29.5% of trials, Subtype 2 for 4.9%, and Subtype 3 for 65.4% of trials. Each group contained both red- and blue-dominant stimuli, but with distinct compositions (**Fig. 2B**): Subtypes 1 and 2 contained predominantly red-dominant trials (86.8% and 61.2%), whereas Subtype 3 contained predominantly blue-dominant trials (67.5%). A small subset of trials formed a low-activation (quiescent) group in which FC-2 activity was effectively absent; we analyze these trials as Subtype 2.

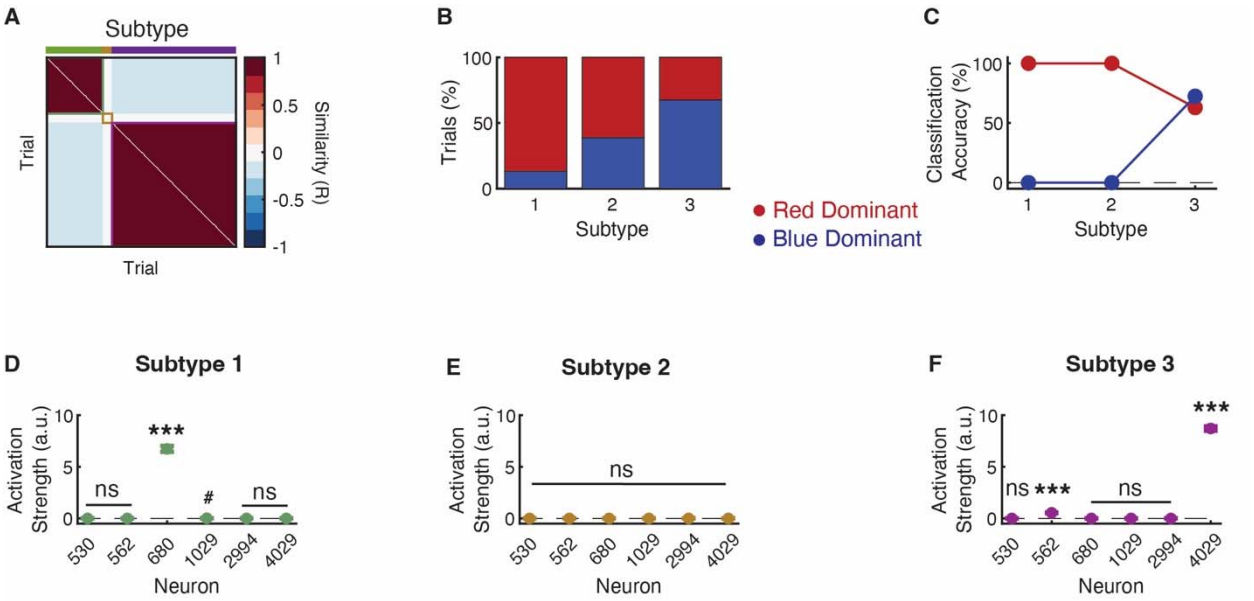

Figure 2. CNN exhibits structured variability during perceptual decision-making task. (A) Modularity-maximization based clustering identified three subtypes of trials based on the activation pattern in layer FC-2. The colored squares correspond to the trials composing each subtype. (B) Composition of trials in each subtype. (C) Classification accuracy per subtype. (D) Activation strength of neurons in Subtype 1. Dot plot shows the average activation value of neurons that were active in at least one trial. Plotted activation value was estimated based only on trials within the subtype (mean ± sem). (E, F) Same as Panel C, but for Subtype 2 and Subtype 3, respectively. Statistical significance was determined using one-sample *t*-test and Bonferroni corrected for multiple comparison. ***$P_{Bonferroni} < 0.001$; $^{\#}P < 0.05$; ns = not significant.

Classification accuracy differed systematically across subtypes (**Fig. 2C**). All red-dominant trials were correctly classified in Subtypes 1 and 2, and 62.9% of red-dominant trials were correctly classified in Subtype 3. In contrast, blue-dominant trials were correctly classified exclusively within Subtype 3.

Inspection of FC-2 activity confirmed distinct activation profiles across subtypes. Across all trials, only six neurons in FC-2 were active at least once. In Subtype 1, two neurons showed significantly greater-than-zero average activation—Neuron-680 (Activation = 6.73 ± 0.28) and Neuron-1029 (Activation = $2.37 \times 10^{\square\square} \pm 2.37 \times 10^{\square\square}$; **Fig. 2D**). In Subtype 2, no measurable activation was observed in any neuron (**Fig. 2E**). In Subtype 3, two neurons showed robust activation—Neuron-562 (Activation = 0.55 ± 0.03) and Neuron-4029 (Activation = 8.7 ± 0.21; Fig. 2F). Trial-level activations are shown in Fig. S1. Together, these results show that deterministic CNN computations can nonetheless express structured trial-level variability in internal activity.

Trial-to-trial variability reflects decision-making under varying stimulus-dependent evidence

The central question is to determine how fluctuations in activity map onto the computations that drive decisions. To address this unknown, we leverage having full access to the network's internal state to directly probe the signal underlying classification. This approach allows us to isolate the components driving decision-making in each subtype.

Classification is based on two signals (logits), Z_{red} and Z_{blue}, estimated as:

$$Z_{red} = W_{red}^{T} \cdot A + C_{red}$$
$$Z_{blue} = W_{blue}^{T} \cdot A + C_{blue}$$

where A is the activation vector in FC-2, W is the corresponding weights, and C is the bias (i.e. constant) term. Z_{red} and Z_{blue} are then converted to probabilities, $P(red)$ and $P(blue)$, using *softmax*. The class, red or blue, to which an image is assigned is based on the highest value using *argmax* between $P(red)$ and $P(blue)$. However, converting probabilities to binary choices does not provide any information on how strongly one option is favored over another. To retain graded information about how strongly one choice is favored, we instead quantified a decision preference (DP) as the difference between red and blue probabilities:

$$DP = P(red) - P(blue)$$

where $DP > 0$ favors red and $DP < 0$ favors blue. From this, we can estimate how stimulus-dependent activity and stimulus-independent bias components shape decision-making. The stimulus-dependent activity component for red- or blue-dominance is represented by $W^{T} \cdot A$. Whereas the stimulus-independent bias component is represented by the bias (i.e. constant) terms with the highest value $(C_{red} - C_{blue})$. The advantage of using DP is that it can be used to determine how evidence and bias shape decision-making. This can be accomplished by setting the bias terms to zero ($C_{red} = C_{blue} = 0$) when estimating Z_{red} and Z_{blue}. In combination, these measures identify when decisions are supported primarily by evidence versus when they rely disproportionately on bias.

First, we focus on red-dominant trials since these could be correctly classified in all three subtypes. As can be observed in Figure 3A, for Subtype 1, the bias terms, C_{red} and C_{blue}, were

not critical for classification because the decision preference favored a classification of red dominance with (DP_{+C} = 0.42 ± 0.006; t(256) = 66.00; $P_{Bonferroni}$ < 0.001) and without the bias terms (DP_{-C} = 0.18 ± 0.007; t(256) = 24.2, $P_{Bonferroni}$ < 0.001; **Fig. 3A**). These results indicate that in Subtype 1 the sensory evidence provided sufficient information favoring one option – red dominant.

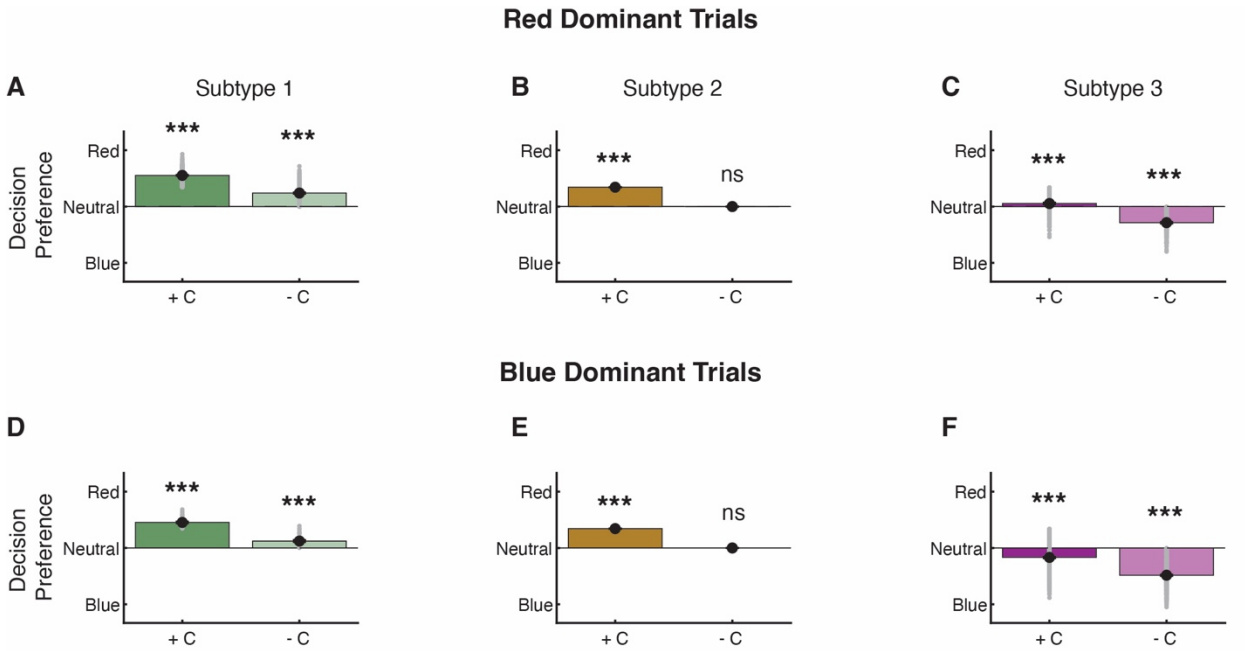

Figure 3. **Trial-level variability reflects decision-making under varying sensory evidence.** (A-C) Decision preference estimated with (+C) and without (-C) bias terms for red dominant trials in Subtype 1, Subtype 2, and Subtype 3, respectively. Bar plot shows the average decision preference value across trials (mean ± sem). Gray dots show individual trials. (D–F) Same as panels A–C, but for blue-dominant trials. Statistical significance was determined using one-sample *t*-test and Bonferroni corrected for multiple comparison. ***$P_{Bonferroni}$ < 0.001; ns = not significant.

For Subtype 2, the bias term was critical in classifying trials as red dominant (DP_{+C} = 0.25 ± 0.0; $P_{Bonferroni}$ < 0.001) because without the bias terms the decision preference dropped to zero as expected given no neurons were active (DP_{-C} = 0.0; t(29) = 0.00, P = 1.00; **Fig. 3B**). Similarly, in Subtype 3, inclusion of the bias term resulted in a subset of trials favoring a red classification (DP_{+C} = 0.04 ± 0.01; t(212) = 3.90, $P_{Bonferroni}$ = 3.87 x 10^{-4}), but without the bias terms the decision preference favored classifying the trials as blue-dominant (DP_{-C} = -0.21 ± 0.01; t(212) = -22.43, $P_{Bonferroni}$ < 0.001), indicating classification based on sensory evidence alone provided conflicting information (**Fig. 3C**).

For blue-dominant trials, Subtype 3 was the primary cluster supporting correct classification. In Subtype 1, stimulus-dependent evidence component was insufficient to classify trials as blue dominant (DP_{-C} = 0.09 ± 0.01; t(38) = 7.58, $P_{Bonferroni}$ = 1.23 x 10^{-8}) and including bias further increased preference toward red (DP_{+C} = 0.34 ± 0.01; t(38) = 32.05, $P_{Bonferroni}$ < 0.001; **Fig. 3D**). Similar patterns were observed for Subtype 2 (**Fig. 3E**). In Subtype 3, the decision preference favored blue under both computations (DP_{-C} = -0.36 ± 0.01; t(441) = 45.00, $P_{Bonferroni}$ < 0.001), but inclusion of bias reduced the magnitude for blue preference, consistent with bias partially counteracting evidence (DP_{+C} = -0.13 ± 0.01; t(441) = -13.66, $P_{Bonferroni}$ < 0.001; **Fig. 3F**). Together, these results suggest that the three subtypes correspond to distinct decision modes: an evidence-dominant mode, a bias-dominant mode when stimulus-driven activity is minimal, and a conflict/near-boundary mode in which bias can partially counteract sensory evidence.

<u>Bias terms preserve performance against noise in the model weights</u>

Noise is a challenge that both biological and artificial systems must contend with. Given that the subtypes reflect distinct balances between evidence and bias, we next asked how bias shapes decisions as internal noise increasingly degrades the evidence pathway. We therefore tested whether bias contributes causally to performance when the evidence pathway is degraded. For each test trial, we added Gaussian noise, $N(0, \sigma)$, either to FC-2 activations or to the classifier weights mapping FC-2 representations to output logits, and quantified accuracy with and without bias.

Noise in the weights was more detrimental to classification accuracy than noise in activations (**Fig. 4A**). Importantly, when noise was applied to the weights, retaining the bias term preserved classification performance relative to a bias-removed computation (**Fig. 4B**). In contrast, when noise was applied to activations, the bias term had minimal impact on accuracy (**Fig. 4C**). These results indicate that bias is not merely descriptive: it can stabilize decisions when the evidence readout itself becomes unreliable.

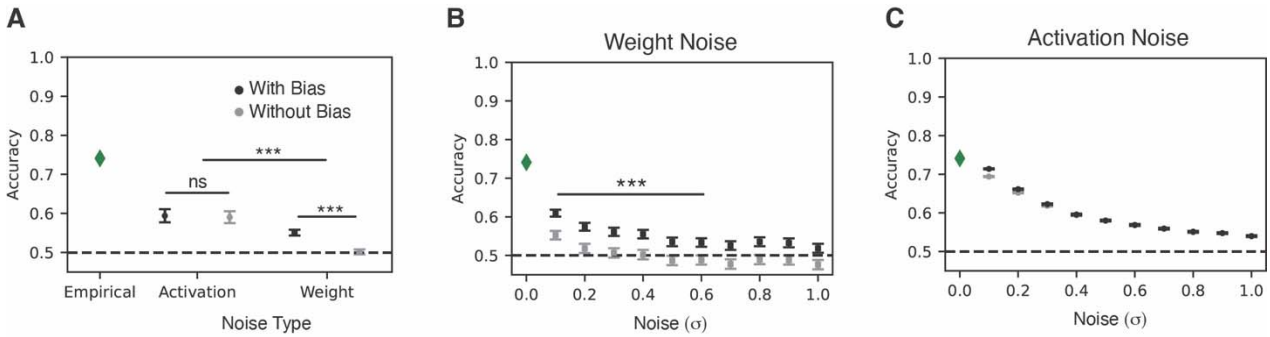

Fig 4. Bias terms preserves performance against noise in the weights. (A) Average classification accuracy after adding noise. Green diamond represents the empirical performance with no added noise ($\sigma = 0$) is shown for reference. Black and gray dots represent the average classification value across all noise levels for with and without bias term in the classification, respectively. The dashed line indicates chance performance (0.5). Error bars show mean ± sem. (B) Classification accuracy for each noise level with added noise to the weights. (C) Same as panel B, but noise added to the neuronal activations. Statistical significance was determined using paired-sample *t*-test. ***P < 0.001; ns = not significant.

Corroborating the findings in additional CNN models

To assess generality across models, we repeated the trial-level clustering and decision-variable decomposition in four additional independently trained CNNs. In all models, trial-level variability in internal activity could be clustered into multiple subtypes (**Figs. S3-S6**), and the bias terms remained a critical component of the decision computation. Thus, the evidence–bias structure observed in the representative model reflects a reproducible feature of this trained CNN class rather than an idiosyncrasy of a single network instance.

Bias throughout the processing hierarchy

Beyond the final classification layer, bias terms pervade the network because every affine transformation—convolutional or fully connected—adds an intercept. Consequently, stimulus-independent offsets can shape the decision variable not only at the readout but also through cumulative contributions across the hierarchy. To quantify where and when decisions are supported primarily by stimulus-dependent activity versus stimulus-independent bias terms, we developed a Bias Dominance Index (BDI).

BDI measures the fraction of the decision margin attributable to the bias term by decomposing the logit margin between a target class and its closest alternative into a stimulus-dependent (evidence) component and a stimulus-independent (bias) component. We computed BDI at successive stages of the network to localize where along the processing hierarchy bias-dominant support emerges.

For the perceptual decision task, BDI was computed on the difference in logit margin between red and blue classes. Across the five independently trained CNNs, bias dominance was not confined to the final output readout (**Fig. 5A**). Instead, multiple convolutional and classifier layers exhibited BDI > 0.50, indicating that bias-like contributions can accumulate across depth and, in many trials, can provide the predominant support for the decision variable.

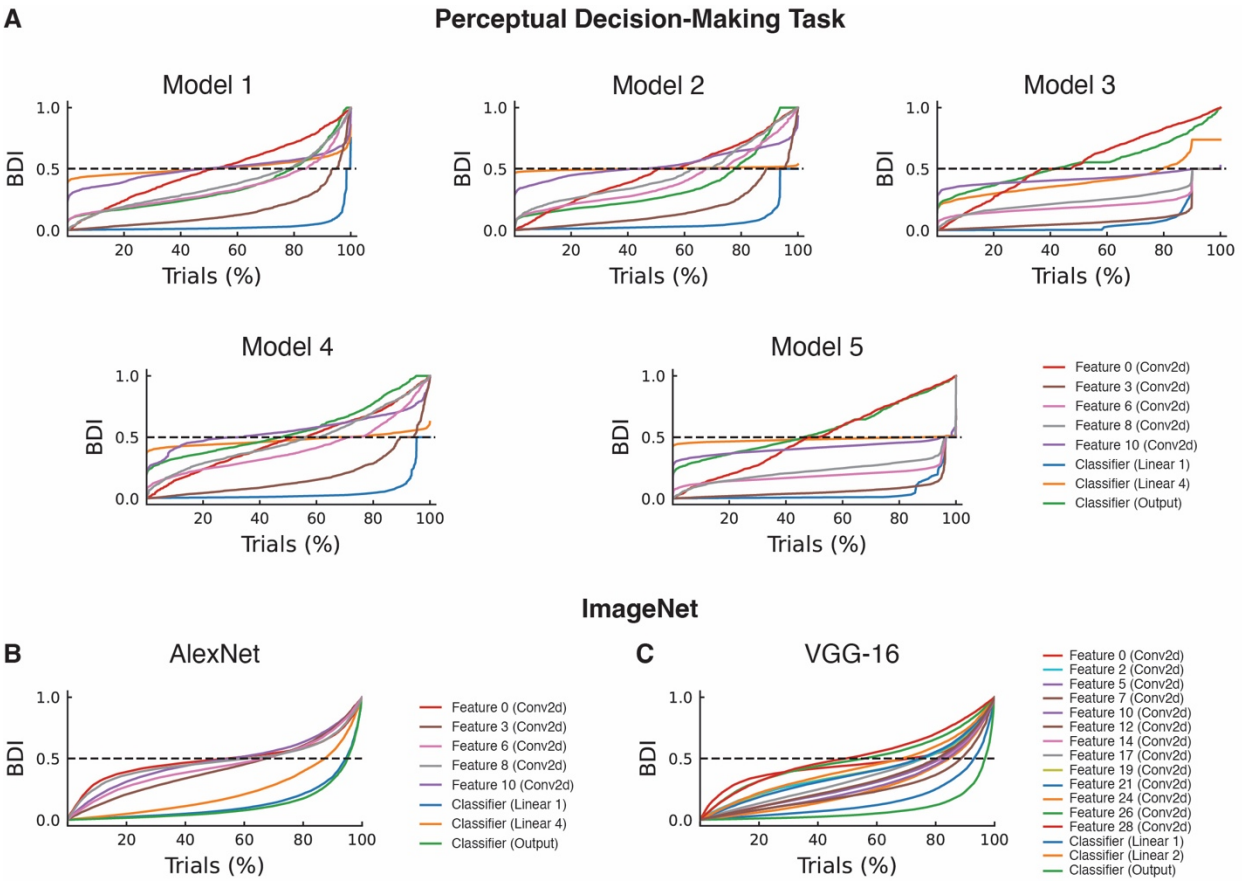

Figure 5. Multi-layer bias dominance in the perceptual decision-making task and ImageNet dataset. (A) Layer-wise influence of bias in perceptual decision-making task. For each of five independently trained CNN models, the Bias Dominance Index (BDI) estimates the fraction of the decision margin attributable to the bias term. The logit margin between a target class and its closest alternative is decomposed into an activation-dependent (evidence) component and a stimulus-independent (bias) component. Colored lines depict trial ranked BDI estimates for layers spanning early convolutional feature blocks to fully connected classifier layers. The dashed line marks BDI = 0.5, the boundary at which the bias term contribution exceeds the activation-dependent (evidence) contribution to the margin. (B) ImageNet evaluation using AlexNet. BDI was computed on the top-1 versus top-2 logit margin for each image (no ground truth required), trials were ranked by BDI, and layer-wise BDI curves are shown as in panel A. (C) Same as panel B, but for VGG-16.

We next tested whether this signature generalizes to natural image classification and other architectures. We applied the same analysis to pretrained AlexNet and VGG-16 on ImageNet data, computing BDI on the top-1 versus top-2 logit margin (a criterion that does not require ground truth labels). Consistent with the perceptual task, both AlexNet (**Fig. 5B**) and VGG-16 (**Fig. 5C**) contained multiple convolutional and classifier layers with BDI > 0.50 indicating that distributed bias-like support is not specific to the dot task or to a single architecture. Together, these results show that bias terms can make a measurable—and sometimes predominant—contribution to decision variables, and that this contribution is expressed across multiple layers of the processing hierarchy.

Discussion

Fluctuations in brain activity are ubiquitous during both simple and cognitively demanding tasks and are often indicative of variation in cognitive processing (17, 18). Here we used a controlled deep-learning model to isolate how variability in internal activity can translate into distinct decision modes. Although the network is deterministic, trial-level fluctuations in its internal representations organized into reproducible cluster subtypes, and, critically, these subtypes differed in how decisions were supported by the decision variable. By decomposing the decision margin into stimulus-dependent evidence and stimulus-independent bias, we show that bias becomes disproportionately influential when evidence is weak, ambiguous, or effectively absent. This bias–evidence decomposition was not merely descriptive, but bias terms preserved performance when the readout was perturbed by noise added to the weights, indicating that stimulus-independent components can stabilize decisions.

A central contribution of this work is to elevate bias from an output-layer “offset” to a distributed computational signal that can accumulate across processing depth. Bias terms are present at every affine stage of deep networks, and the layer-resolved Bias Dominance Index (BDI) quantified where along the processing hierarchy bias-like contributions dominate the decision margin. Across independently trained models on the perceptual task, and across pretrained architectures on natural image classification, BDI revealed that bias-dominant support is not confined to the final readout. Instead, multiple convolutional and classifier stages can contribute stimulus-independent support to the decision variable, consistent with the idea that bias can be progressively shaped by the hierarchy rather than injected solely at the end (19). This provides a generalizable diagnostic for when predictions are likely to be bias-weighted versus evidence-driven, independent of a specific task, architecture, or output dimensionality.

These results also clarify how distinct decision modes (20) and strategies (21) can arise without invoking stochastic state switches or discrete latent states. In our setting, variability is driven by controlled differences in sensory input, yet it produced separable clusters in internal activity and systematic shifts in the balance between evidence and bias. Importantly, this does not imply that brain variability is purely bottom-up (9). Instead, it suggests that bottom-up dynamics may be sufficient to generate distinct computational modes, which can then be further shaped by top-down attention (7), and recurrent dynamics (8).

Moreover, the states observed in the CNN suggest that the multiple patterns of activation identified in humans performing the same perceptual decision-making task may likewise reflect shifts in the balance between stimulus-dependent evidence and stimulus-independent bias. In particular, these results provide insight pertaining to the role of the Default Mode Network in decision-making. In the brain, increased engagement of the Default Mode Network is associated with a state in which external inputs contribute little to brain activity (22). Analogously, in the CNN, Subtype 2 corresponded to a state in which the stimulus contributed little to activity. Nevertheless, decisions could still be generated both during Default Mode Network-dominant periods in humans and for trials in Subtype 2 of the CNN. We therefore speculate that engagement of the Default Mode Network may be associated with decisions that rely more on bias than on sensory evidence, paralleling decisions driven by the bias terms observed in the CNN.

Stimulus-independent bias is typically formalized as a criterion shift in signal detection theory (23), shaped by expectations, priors, payoffs, or learned asymmetries (24). In humans, bias-dominant decisions has been shown to play a role in decision-making (25) and can be beneficial in cases where one alternative is preferred over another (26). Moreover, bias-dominant decisions may play a similar role as in the CNN and may not simply be a confounding factor (27, 28). How bias-dominant decisions arises in artificial neural networks and the brain is an open question, but prior work suggests that bias may reflect competition between different components of the network (29).

Beyond neuroscience, these findings have practical implications for interpreting and auditing modern neural networks. Because bias-like support can dominate across layers, networks may produce confident-looking predictions even when stimulus-dependent evidence is weak. This is

consistent with prior work showing that the bias terms can make substantial contributions to prediction accuracy. In certain settings, it has been shown that using only the bias terms can achieve roughly 30–40% of full-model performance and backpropagating only the bias was sufficient to reconstruct an image (30). BDI offers a principled tool to flag such modes by quantifying the relative contribution of bias to the decision margin across depth. More generally, comparing evidence-only margins to full margins provides a simple diagnostic to stratify predictions into evidence-driven versus bias-weighted modes, which could inform uncertainty estimation, selective prediction, or calibration strategies.

This work contains several limitations. First, our analyses use feedforward CNNs and future work should extend these analyses to architectures with recurrence, attention, and explicit state dynamics where internally generated fluctuations can be probed. Second, low-activation clusters in rectified networks can reflect nonlinear gating and representational sparsity; this is partially, mitigated by findings that bias can dominate in AlexNet and VGG-16 when classifying images from the ImageNet dataset, nonetheless, establishing the robustness of quiescent states should be established in other perceptual decision-making tasks. Finally, the links between CNN activity and brain activity is only analogous and future work is needed to test, for instance, whether DMN-dominant periods reflect bias-driven decision.

Take together, our findings outline a unifying framework on the relationship between fluctuation in activity and decision-making: both biological and artificial systems organize trial-to-trial variability into a small set of discrete states. Understanding how systems make decisions as evidence waxes and wanes could provide a foundation for studying disrupted decision-making where abnormal fluctuations may reflect impairment. Overall, variation in activity over time is not a byproduct of noise but a feature of adaptive decision-making, enabling flexible responses under varying levels of evidence and uncertainty.

Materials and Methods

Task description

The CNN was trained to perform a perceptual decision-making task, identifying whether red or blue dots were more frequent in a cloud of randomly distributed dots. Specifically, a CNN was trained to perform a modified version of a perceptual decision-making task in which we have previously identified multiple patterns of activation(4). Each stimulus consisted of either 90 red- or 100 blue-colored dots, presented within an imaginary circular aperture. Dot positions were

randomized for every trial to minimize low-level visual regularities. In total, 1,000 novel stimuli were created for testing with equal number of red- or blue-dominant trials. All images were generated using Python.

CNN model, training and testing

To unravel the relationship between fluctuations in activity and decision-making, we opted to use a CNN model because it could be trained to perform perceptual decision-making tasks and because it provides complete access to the internal states. We used a CNN based on the AlexNet architecture, modified for binary classification (red- versus blue-dominant images). CNN architecture is shown in Table S1. The weights associated with the convolutional layers were fixed corresponding to pretrained on ImageNet (1K classes), as provided by the PyTorch library (https://pytorch.org/). Only the fully connected layers were trainable, allowing the network to adapt the decision stage while maintaining a stable feature-extraction stage.

The CNN was trained on randomly generated trials of the perceptual decision-making task. Training was conducted in five batches of randomly chosen trials. Model optimization was performed using stochastic gradient descent as implemented in PyTorch. After training, the CNN was evaluated on 1,000 unique trials not included in the training set. Accuracy was assessed separately for red-dominant and blue-dominant images, as well as across all trials. In total five different CNN models were trained and tested. All training and testing was conducted using Google Colab on an NVIDIA A100 GPU.

ImageNet dataset

To assess the generality of the analysis beyond the perceptual decision-making task, we evaluated pretrained convolutional neural networks on images from the Tiny ImageNet dataset (Tiny ImageNet-200), a 200-class subset of ImageNet(31). We analyzed all 10,000 images from the held-out split test set for unlabeled analyses. Images were resized and center-cropped to 224×224 and normalized using the standard ImageNet-1K preprocessing associated with the pretrained model weights. Network outputs were taken as the 1000-way ImageNet-1K logits.

Activation of the penultimate CNN layer

For each trial in the test set, we extracted the activation patterns from the penultimate layer (FC-2). For each trial, the activation vector A is the set of activations across all neurons in the FC-2 layer elicited by the stimulus. A neuron was considered active if the value in the activation vector,

A, in the penultimate layer was greater than zero on at least on trial. Operationally, one presents a stimulus image to the trained CNN, performs a forward pass through the network up to FC-2, and then records the FC-2 outputs for that trial as a vector of neuron activations.

Modularity-maximization based clustering

The activation patterns formed the basis for subsequent analyses. To identify structure in variability, we applied clustering to FC-2 activations, grouping trials with similar patterns of activity into discrete subtypes. A trial-by-trial similarity matrix was created using the Pearson Product Correlation based on the activation. To minimize noise, the Pearson correlation was computed using only neurons that were activate on at least one trial.

Clustering of the similarity matrix was conducted using modularity-maximization (32). Modularity-maximization has been primarily used to detect clusters of nodes or regions in brain networks, but modularity-maximization can be used to cluster other types of data (33). We have previously used modularity maximization to identify multiple activation patterns in brain activity from fMRI data (4) and from EEG data (34, 35). Moreover, we have established a framework that utilizes modularity maximization to identify robust clusters in neuroimaging data which we utilized in the current analysis (16). Furthermore, unlike other clustering methods, modularity-maximization does not require *a priori* specifying the number of clusters minimizing user induced bias.

The clustering resolution in modularity-maximization is specified using the resolution parameter (γ). In the current analysis, $\gamma = 1$. The clustering results using modularity-maximization are not deterministic, and the results can depend on the specific random seeds. This limitation can be overcome using consensus clustering to identify stable clusters out of a set of partitions (36). In our previous work of clustering single trials, we had found that 100 iterations were sufficient to identify stable clusters (37). Specifically, consensus clustering identifies a single representative partition from the set of 100 iterations. This process involves the creation of a thresholded association matrix which describes how often nodes were placed in the same cluster. The representative partition is then obtained by using a Generalized Louvain algorithm to identify clusters within the thresholded association matrix. Modularity-maximization was implemented with the Generalized Louvain algorithm part of the Brain Connectivity Toolbox (38).

The interplay between sensory evidence and choice-bias in the classifier

For each trial, the CNN produces two class scores (logits), Z_{red} and Z_{blue}, from the activations of the second fully connected layer (FC-2):

$$Z_{red} = W_{red}^{T} \cdot A + C_{red}$$
$$Z_{blue} = W_{blue}^{T} \cdot A + C_{blue}$$

where A is the FC-2 activation vector for that trial, W_{red} and W_{blue} are the corresponding weight vectors, and C_{red} and C_{blue} are the bias (i.e. constant) terms. Logits are converted to class probabilities via the *softmax* function:

$$P(red) = \frac{e^{Z_{red}}}{e^{Z_{red}} + e^{Z_{blue}}}$$

$$P(blue) = \frac{e^{Z_{blue}}}{e^{Z_{red}} + e^{Z_{blue}}}$$

The assigned class (red vs. blue) is given by $\text{argmax}\,\{P(\text{red}), P(\text{blue})\}$. To quantify graded choice tendencies beyond the binary decision, we computed decision preference (DP) as:

$$DP = P(red) - P(blue)$$

when $DP > 0$ indicates a preference for red and $DP < 0$ indicates a preference for blue. DP was computed with the bias terms intact (DP_{+C}) and with bias removed (DP_{-C}). by setting intercepts to zero ($C_{red} = C_{blue} = 0$). In this decomposition, the dot-product terms $W_{red}^{T} \cdot A$ and $W_{blue}^{T} \cdot A$ capture the stimulus-dependent evidence for each class, whereas the bias terms capture an input-independent bias that can shift the effective decision boundary.

<u>Noise perturbations of stimulus representations and readout</u>

For each trial in the test set, noise was applied either to the FC-2 activation pattern or to the readout weights, thereby perturbing the stimulus representation or its mapping to the output, respectively. Activation noise was modeled by adding Gaussian noise, $\mathcal{N}(0, \sigma)$, to the neuronal activations in FC-2, which reflect the network's internal representation of the stimulus. Weight noise was modeled by adding Gaussian noise, $\mathcal{N}(0, \sigma)$, to the weights mapping FC-2 activations to the classification output. The standard deviation of the noise (σ) ranged from 0.1 to 1.0, and classification accuracy was estimated from 1,000 simulations at each noise level with and without the bias terms.

Bias dominance index

Bias Dominance Index (BDI) is used to quantify the relative contribution of stimulus-dependent versus stimulus-independent components of the decision variable on a pairwise logit margin between a designated target class t and alternative class r. BDI builds on prior work that aims to attribute network behavior to specific underlying factors (30, 39, 40).

For a classifier with N output classes, logits are given by a linear readout from an internal activation vector A:

$$z_k = W_k^\top A + C_k, k \in \{1, \dots, N\}$$

where W_k and C_k denote the weights and bias (intercept) for class k. For each trial, we define the logit margin between target class t and alternative class r as $\Delta z_{t,r} = z_t - z_r$, which decomposes into an evidence-dependent term (Δz_{evid}) and a bias term (Δz_{bias}):

$$\Delta z_{\text{evid}} = (W_t - W_r)^\top A$$
$$\Delta z_{\text{bias}} = b_t - b_r$$

We then define the bias dominance index as:

$$\text{BDI}_{t,r} = \frac{|\ \Delta z_{\text{bias}}\ |}{|\ \Delta z_{\text{bias}}\ | + |\ \Delta z_{\text{evid}}\ |}$$

By construction, $\text{BDI}_{(t,r)} \in [0,1]$. Values near 0 indicate trials dominated by the stimulus-dependent component, whereas values near 1 indicate trials dominated by stimulus-independent component (i.e. bias/intercept). Values greater than 0.5 indicate that the stimulus-independent component is larger in magnitude than the stimulus-dependent component for that margin.

To extend BDI beyond the final readout, an analogous bias–evidence decomposition at multiple stages of the network is computed. For each affine transformation at layer ℓ (convolution or fully connected), its pre-nonlinearity output is $u^{(\ell)} = W^{(\ell)} A^{(\ell-1)} + C^{(\ell)}$ and is separated into a stimulus-dependent component $u_{\text{evid}}^{(\ell)} = W^{(\ell)} A^{(\ell-1)}$ and a stimulus-independent component $u_{\text{bias}}^{(\ell)} = C^{(\ell)}$ (broadcast across spatial positions for convolutional layers).

For a given trial-wise margin $\Delta z_{t,r}$, we quantify the layer's contribution using a gradient-based attribution:

$$C_{\text{evid}}^{(\ell)} = \left\langle \frac{\partial \Delta z_{t,r}}{\partial u^{(\ell)}}, u_{\text{evid}}^{(\ell)} \right\rangle,$$

$$C_{\text{bias}}^{(\ell)} = \left\langle \frac{\partial \Delta z_{t,r}}{\partial u^{(\ell)}}, u_{\text{bias}}^{(\ell)} \right\rangle$$

and define a layer-specific index as:

$$\text{BDI}^{(\ell)} = \frac{|\, C_{\text{bias}}^{(\ell)} \,|}{(|\, C_{\text{bias}}^{(\ell)} \,| + |\, C_{\text{evid}}^{(\ell)} \,|)}$$

Using the final logit margin $\Delta z_{t,r}$ anchors BDI to the model's true decision variable and avoids probe-dependent assumptions.

The most generalizable form of the analysis is to estimate BDI on each trial between the primary and secondary choices since this does not require a ground truth and can be applied to cases with multiple classes. Specifically, the primary choice is the predicted class $t = argmax_{k=t}(z_k)$ and the class associated with the second-largest logit $r = argmax_{k \neq t}(z_k)$. This defines BDI on the smallest "winning margin" separating the model's top prediction from its nearest alternative. If the ground truth is known, then the target, t, is logit from the ground-truth class and the alternative, r, is the highest-logit from the incorrect class. This yields a label-anchored measure of whether correct classification is supported primarily by stimulus-dependent activity or by bias. In the special case of two classes, competitor selection is fixed by the task. In the perceptual decision-making task, t corresponded to the red class and r to the blue class (or vice versa), therefore BDI is computed on the difference in logit margin between red and blue classes.

Data and code availability

Trained CNN models are available at https://osf.io/efnq6/. ImageNet data can be download at http://cs231n.stanford.edu/tiny-imagenet-200.zip. The analysis was based on a combination of publicly available toolboxes, analysis specific scripts. Specifically, the CNNs were based on AlexNet architecture from the PyTorch library (https://pytorch.org/). Formal analysis was conducted in MATLAB 2021b. Trained AlexNet and VGG-16 can be downloaded from PyTorch library (https://pytorch.org/). Trained CNN models, scripts for training and formal analysis are

available at https://osf.io/efnq6/. Clustering analysis was based on the Community Louvain part of the Brain Connectivity Toolbox (https://sites.google.com/site/bctnet/). Consensus clustering was determined using *consensus_iterative.m* (http://commdetect.weebly.com/).

References

1. R. L. T. Goris, J. A. Movshon, E. P. Simoncelli, Partitioning neuronal variability. *Nature Neuroscience* **17**, 858–865 (2014).

2. A. Arieli, A. Sterkin, A. Grinvald, A. Aertsen, Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses. *Science* **273**, 1868–1871 (1996).

3. L. Waschke, N. A. Kloosterman, J. Obleser, D. D. Garrett, Behavior needs neural variability. *Neuron* **109**, 751–766 (2021).

4. J. Nakuci, *et al.*, Multiple brain activation patterns for the same perceptual decision-making task. *Nature Communications* **16**, 1785 (2025).

5. K. Christoff, Z. C. Irving, K. C. R. Fox, R. N. Spreng, J. R. Andrews-Hanna, Mind-wandering as spontaneous thought: a dynamic framework. *Nature Reviews Neuroscience* **17**, 718–731 (2016).

6. M. F. Mason, *et al.*, Wandering Minds: The Default Network and Stimulus-Independent Thought. *Science* **315**, 393–395 (2007).

7. N. Al-Aidroos, C. P. Said, N. B. Turk-Browne, Top-down attention switches coupling between low-level and high-level areas of human visual cortex. *Proceedings of the National Academy of Sciences* **109**, 14675–14680 (2012).

8. S. Peron, *et al.*, Recurrent interactions in local cortical circuits. *Nature* **579**, 256–259 (2020).

9. M. Corbetta, G. L. Shulman, Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience* **3**, 201–215 (2002).

10. R. Rajalingham, *et al.*, Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience* **38**, 7255–7269 (2018).

11. D. L. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience* **19**, 356–365 (2016).

12. N. Kriegeskorte, R. A. Kievit, Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences* **17**, 401–412 (2013).

13. R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific reports* **6**, 27755 (2016).

14. M. Eickenberg, A. Gramfort, G. Varoquaux, B. Thirion, Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage* **152**, 184–194 (2017).

15. S.-M. Khaligh-Razavi, N. Kriegeskorte, Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology* **10**, e1003915 (2014).

16. J. Nakuci, D. Rahnev, A Practical Guide to Identifying Robust Clusters in Neuroimaging Data. *Human Brain Mapping* **46**, e70330 (2025).

17. A. A. Faisal, L. P. J. Selen, D. M. Wolpert, Noise in the nervous system. *Nature Reviews Neuroscience* **9**, 292–303 (2008).

18. D. D. Garrett, *et al.*, Moment-to-moment brain signal variability: A next frontier in human brain mapping? *Neuroscience & Biobehavioral Reviews* **37**, 610–624 (2013).

19. L. Serrano-Fernández, R. Romo, N. Parga, Early emergence of perceptual biases in the secondary somatosensory cortex. *Proceedings of the National Academy of Sciences* **122**, e2519255122 (2025).

20. V. Weilnhammer, H. Stuke, K. Standvoss, P. Sterzer, Sensory processing in humans and mice fluctuates between external and internal modes. *PLoS Biology* **21**, e3002410 (2023).

21. Z. C. Ashwood, *et al.*, Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience* **25**, 201–212 (2022).

22. M. D. Greicius, B. Krasnow, A. L. Reiss, V. Menon, Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* (2003). https://doi.org/10.1073/pnas.0135058100.

23. D. Linares, D. Aguilar-Lleyda, J. López-Moliner, Decoupling sensory from decisional choice biases in perceptual decision making. *eLife* **8**, e43994 (2019).

24. S. A. Klein, Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & psychophysics* **63**, 1421–1455 (2001).

25. J. I. Gold, L. Ding, How mechanisms of perceptual decision-making affect the psychometric function. *Progress in Neurobiology* **103**, 98–114 (2013).

26. M. J. Mulder, E.-J. Wagenmakers, R. Ratcliff, W. Boekel, B. U. Forstmann, Bias in the brain: a diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience* **32**, 2335–2343 (2012).

27. R. Akaishi, K. Umeda, A. Nagase, K. Sakai, Autonomous mechanism of internal choice estimate underlies decision inertia. *Neuron* **81**, 195–206 (2014).

28. A. E. Urai, A. Braun, T. H. Donner, Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias. *Nature communications* **8**, 14637 (2017).

29. L. Lebovich, R. Darshan, Y. Lavi, D. Hansel, Y. Loewenstein, Idiosyncratic choice bias naturally emerges from intrinsic stochasticity in neuronal dynamics. *Nature human behaviour* **3**, 1190–1202 (2019).

30. S. Wang, T. Zhou, J. Bilmes, Bias also matters: Bias attribution for deep neural network explanation in (PMLR, 2019), pp. 6659–6667.

31. Y. Le, X. Yang, Tiny imagenet visual recognition challenge. *CS 231N* **7**, 3 (2015).

32. P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, J. P. Onnela, Community structure in time-dependent, multiscale, and multiplex networks. *Science* **328**, 876–878 (2010).

33. F. Zamani Esfahlani, *et al.*, Modularity maximization as a flexible and generic framework for brain network exploratory analysis. *NeuroImage* **244**, 118607 (2021).

34. J. Nakuci, T. J. Covey, J. L. Shucard, D. W. Shucard, S. F. Muldoon, Single trial variability in neural activity during a working memory task reveals multiple distinct information processing sequences. *NeuroImage* **269**, 119895 (2023).

35. J. Nakuci, J. Samaha, D. Rahnev, Brain signatures indexing variation in internal processing during perceptual decision-making. *iScience* **26**, 107750 (2023).

36. A. Lancichinetti, S. Fortunato, Consensus clustering in complex networks. *Sci. Rep.* **2** (2012).

37. D. S. Bassett, *et al.*, Robust detection of dynamic community structure in networks. *Chaos* **23** (2013).

38. M. Rubinov, O. Sporns, Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage* **52**, 1059–1069 (2010).

39. M. Sundararajan, A. Taly, Q. Yan, Axiomatic Attribution for Deep Networks in *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research., D. Precup, Y. W. Teh, Eds. (PMLR, 2017), pp. 3319–3328.

40. G. Erion, J. D. Janizek, P. Sturmfels, S. M. Lundberg, S.-I. Lee, Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence* **3**, 620–631 (2021).