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

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

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

Keywords: Convolutional neural networks, trial-to-trial variability, perceptual decision-making, sensory evidence, choice-bias

Abstract

Even when stimuli and task demands are held constant, brain activity fluctuates markedly across trials, yet it remains unknown how these fluctuations relate to decision-making and behavior. Here we address this gap in knowledge using a convolutional neural network (CNN) trained on a perceptual decision-making task. Applying an established analytical framework, we show that CNN activity exhibits trial-level variability reflecting findings from human neuroimaging. Notably, the CNN contained a state characterized by minimal stimulus-driven activity, paralleling periods of DMN engagement observed during the same task in humans. Examination of the network's internal state revealed: (i) decisions dominated by strong sensory evidence, (ii) decisions under low discriminability with weak or ambiguous evidence, and (iii) decisions in which sensory evidence was opposed by bias. Moreover, bias stabilized decisions and improved performance when sensory evidence was weak or conflicting. Together, these findings show that variability in CNN activity can serve as a model for understanding how the brain transforms changes in internal states into adaptive behavior, providing a bridge between fluctuations in activity, decision-making, and behavior.

Significance Statement

Brain activity varies substantially from trial to trial, but the role of this variability remains unclear. Using CNNs trained on a perceptual decision-making task, we show that variability in activity organizes into discrete states reflecting evidence-, uncertainty-, and bias-driven decisions. By shifting decision boundaries, CNNs maintain accurate performance even under weak or conflicting evidence. These findings position CNNs as a tractable framework for probing how stimulus driven fluctuations in internal states shape decision-making.

Main Text

Introduction

Brain activity exhibits pervasive trial-to-trial variability despite consistent sensory inputs and task demands (1, 2). Although often dismissed as noise in decision-making models, growing evidence indicates that fluctuations play a critical role in adapting behavior to the demands of a changing environment (3). However, it is not well understood how these fluctuations map unto decision-making and behavior.

In fact, during a perceptual decision-making task, brain regions can exhibit stark variation in activity, yet these fluctuations can have minimal impact on behavioral performance (4). For instance, on a subset of trials the Default Mode Network (DMN) was activated — brain regions commonly associated with mind-wandering (5) and being off task (6) — but accuracy, reaction times, and confidence were comparable to when the DMN was inactive, and regions linked to focused attention were engaged (**Fig. 1A**). The finding that DMN activation does not necessarily impair behavioral performance suggests that DMN engagement may play a functional role in decision-making rather than reflecting mind-wandering or disengagement. Moreover, this dissociation underscores the need to determine how variation in activity maps onto decision-making and behavior.

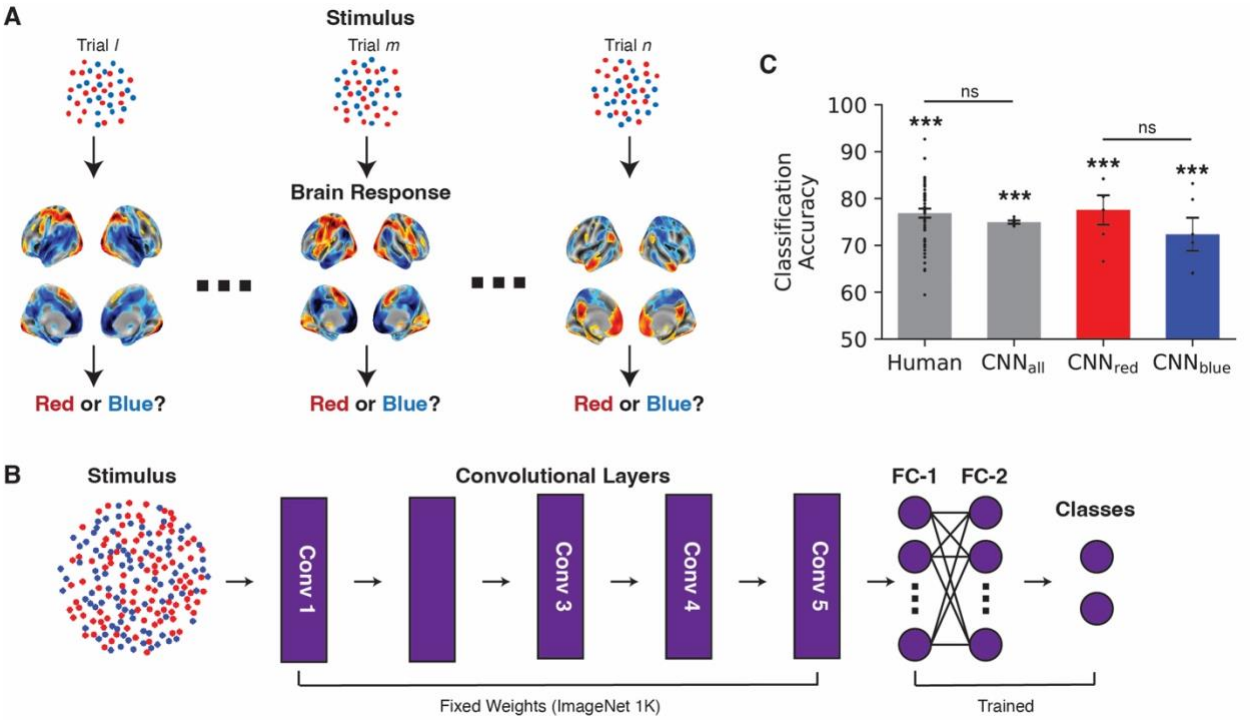

Figure 1. Multiple patterns of activation and CNN implementation of a perceptual decision-making task. (A) Canonical multiple activation patterns identified during a perceptual decision-making task. Subjects were tasked with identifying the dominant color in an image consisting of red and blue dots. 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. ***P < 0.001, **P < 0.01, ns = not significant.

Fluctuations in brain activity have many sources such as top-down processes reflecting changes in attention (7), recurrent connectivity in local circuits (8), or bottom-up processes driven by the stimulus (9). Moreover, measurements of brain activity are inherently noisy. Collectively, these factors make it difficult to determine the primary driver of observed fluctuation in brain activity and how these fluctuations map unto decision-making processes.

Unraveling the relationship between neural activity and decision-making requires a system that provides complete access to the internal state. Convolutional neural networks (CNNs) are such a system, enabling direct investigation of internal states. Trained on objectives comparable to those of the brain, CNNs have been used to gain insight into object recognition (10, 11), representation (12) and processing in the visual cortex (13–15). Building on this foundation, we use a CNN to determine how decision-making changes with fluctuation in activity. Moreover, in CNNs, trial-to-trial variability arises purely from the stimulus providing a controlled setting to isolates the contribution of bottom-up processes in shaping neuronal activity and from which other sources of variability can be investigated.

To anticipate our findings, first we show that during a perceptual decision-making task, the CNN exhibited multiple states reflecting observations in human neuroimaging data. Moreover, the CNN contained a state in which the stimulus contributed little to activity most closely corresponds to a default-like state, analogous to periods of increased DMN engagement observed during the same task in humans. Inspecting the internal state, we found that the trial-level activation patterns corresponded to decisions made when sensory evidence strongly favored one option. A second state reflected decisions under low discriminability, where sensory evidence provided little or no basis for distinguishing between options. A third state reflected bias-driven decisions when sensory evidence provided conflicting information. Collectively, these results suggest that trial-to-trial variation in activity reflects evidence-, uncertainty-, and bias-driven decision-making.

Results

Implementation of a perceptual decision-making task in a CNN

The brain exhibits marked fluctuations across trials, despite constant stimuli and task demands. However, it is not well understood how these fluctuations affect decision-making processes and behavior. To address this gap in knowledge, we trained a CNN to perform a perceptual decision-making task and examined trial-level neural activity. The CNN performed a modified version of a perceptual decision-making task in which we have previously identified multiple activation patterns (4). Specifically, a CNN

was tasked to determine the dominant color in an image containing red and blue dots (**Fig. 1B**). In total, we trained and tested five different CNN models.

Across models, CNNs performed significantly above chance in classifying the dominant color (Accuracy = 74.96 ± 0.38%; $t(4) = 65.41$, $P = 3.27 \times 10^{-7}$) with no significant difference in performance compared to humans (Accuracy = 76.87 ± 0.95%, $t = 1.84$, $P = 0.07$; **Fig 1C**). Moreover, CNNs performed significantly above chance in classifying individual red (Accuracy = 77.57 ± 3.48%; $t(4) = 7.91$, $P = 1.38 \times 10^{-3}$) and blue dominant trials (Accuracy = 72.36 ± 3.95%; $t(4) = 5.65$, $P = 0.004$) and with comparable accuracy between dominant colors ($t(4) = 0.70$, $P = 0.52$), confirming that the CNN was able to perform the task without any preference for a specific color.

<u>CNN contains multiple distinct but stable subtypes of trials during perceptual decision-making</u>

For the analysis, we focus on the activity of neurons in the penultimate layer, FC-2 (**Fig. 1B**). The activity pattern in FC-2 encodes a transformed representation of the input image that is used by the classifier to determine the dominant color (i.e. red or blue). Therefore, trial-to-trial variation in FC-2 directly influences the decision output.

For the analysis, we used an established framework based on modularity-maximization to determine if a CNN exhibits multiple activation patterns similar to what is observed in humans (16). Briefly, the approach groups individual trials based on the similarity in neuronal activation estimated using Pearson correlation. To minimize noise, Pearson correlation was computed using only neurons that were activate on at least one trial.

Focusing on a representative CNN model (Model-4), modularity-maximization based clustering identified three groups of trials, or subtypes (**Fig. 2A**). Subtype 1 accounted for 29.6% of trials, Subtype 2 for 4.9%, and Subtype 3 for 65.5% of trials. All three subtypes contained both red- and blue-dominant trials, but their composition differed (**Fig. 2B**). Subtypes 1 and 2 were composed primarily of red-dominant trials (86.8% and 61.2%, respectively), whereas Subtype 3 contained primarily blue-dominant trials (67.5%). Trials associated with Subtype 2 were initially assigned to their own individual cluster because in these trials all neurons were inactive, therefore we grouped them to form Subtype 2.

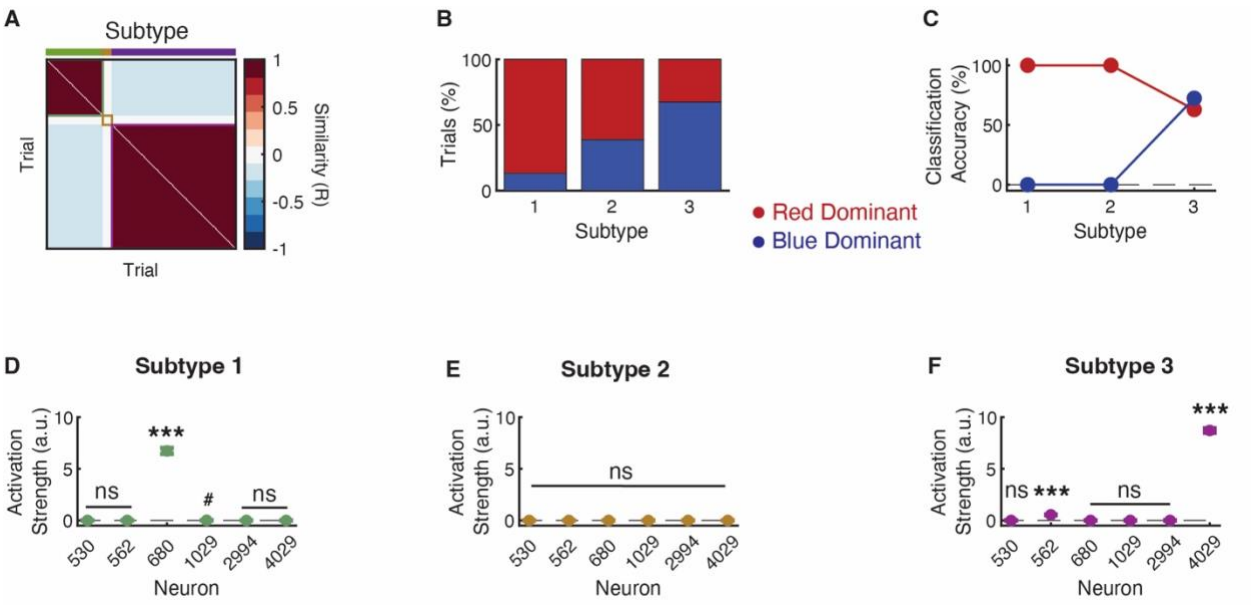

Figure 2. CNN exhibits structured variability during perceptual decision-making task. (A) Modularity-maximization based clustering identified three subtypes of trials based on the activation pattern in layer FC-2. The colored squares correspond to the trials composing each subtype. (B) Composition of 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.

In Subtype 1 and Subtype 2, all red-dominant trials were correctly classified (**Fig. 2C**). Even in Subtype 3, 62.9% of red-dominant trials were correctly classified. On the other hand, blue-dominant trials were exclusively correctly classified in Subtype 3.

Turning to the activity patterns of the neurons driving the clustering, in total, 6 neurons were active in at least one trial, and qualitatively distinct patterns of activity can be observed. In Subtype 1, a significantly greater than zero average activation was observed in Neuron-680 (Activation = 6.73 ± 0.28; $t(295) = 23.84$, $P_{Bonferroni} < 0.001$) and Neuron-1029 (Activation = $2.37 \times 10^{-5} \pm 2.37 \times 10^{-5}$; $t(295) = 2.05$, $P = 0.04$; **Fig. 2D**). By contrast, in Subtype 2, no measurable activation was observed in any neuron (**Fig. 2E**). In Subtype 3, two neurons exhibited an average activation that was significantly greater than zero – Neuron-562 (Activation = 0.55 ± 0.03; $t(654) = 19.10$, $P_{Bonferroni} < 0.001$) and Neuron-4029 (Activation = 8.7 ± 0.21; $t(654) = 42.37$, $P_{Bonferroni} < 0.001$; **Fig. 2F**). Trial-level activations are plotted in Figure S1. These findings confirm that a CNN can exhibit multiple patterns of activation similar to what is observed in humans.

<u>Trial-to-trial variability reflects decision-making under varying sensory evidence</u>

Importantly, the key challenge is to determine how fluctuations in activity shape decision-making. 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:

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

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 preserve this information, we instead estimate the decision preference (DP) which is a measure of how strongly one option is preferred over another as:

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

when $DP > 0$ indicates a preference for red and $DP < 0$ indicates a preference for blue. From this, we can estimate how *sensory evidence* and *choice-bias* shape decision-making. The strength of *sensory evidence* for red- or blue-dominance is represented in their respective $A \cdot W$ component and *choice-bias* by the bias (i.e. constant) term with the highest value. The advantage of using DP is that it can be used to determine how the strength of sensory evidence and bias affects decision-making in each of the subtypes. This can be accomplished setting the bias terms to zero ($C_{\text{red}} = C_{\text{blue}} = 0$) when estimating Z_{red} and Z_{blue}. In combination, this provides a means to determine how activation patterns are linked to specific factors (sensory evidence vs choice-bias) affecting classification.

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 = 0.42 ± 0.006; t(256) = 66.00; $P_{\text{Bonferroni}} < 0.001$) and without the bias terms (DP = 0.18 ± 0.007; t(256) = 24.2, $P_{\text{Bonferroni}} < 0.001$; **Fig. 3A**). These results indicate that in Subtype 1 the sensory evidence provided sufficient information favoring one option – red dominant.

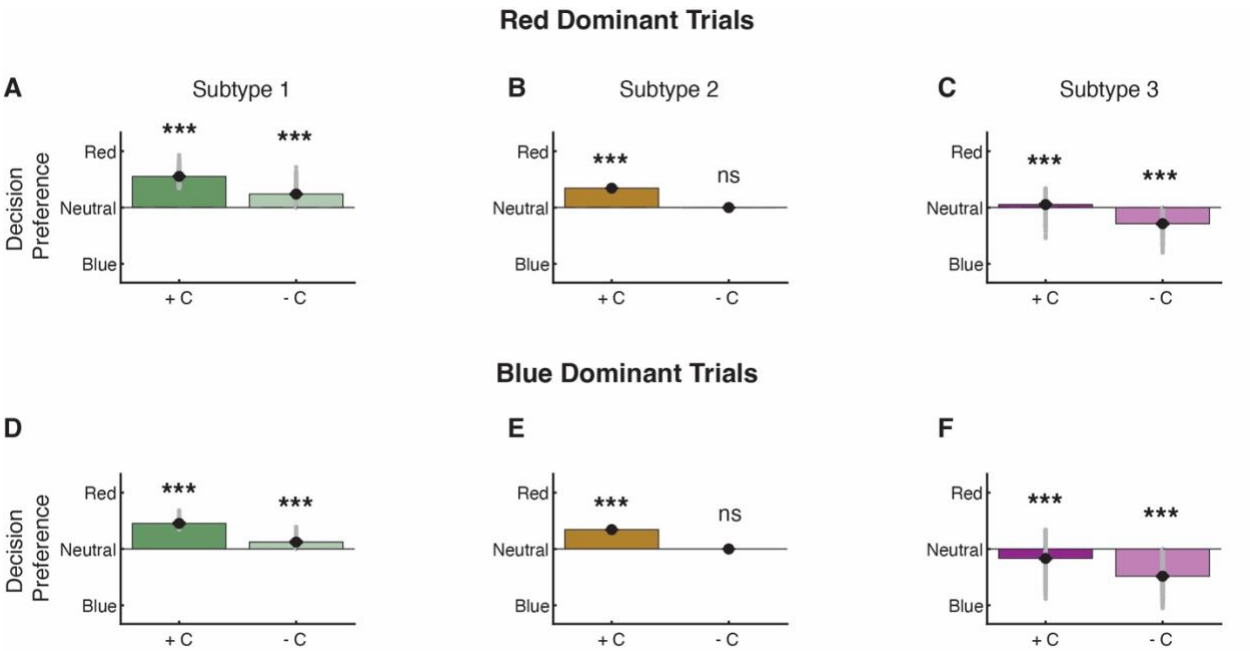

Figure 3. **Trial-level variability reflects decision-making under varying sensory evidence.** (A-C) Decision preference estimated with (+) and without (-) bias terms, C, for red dominant trials in Subtype 1, Subtype 2, and Subtype 3, respectively. Bar plot shows the average decision preference value 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 = 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 = 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 = 0.04 ± 0.01; t(212) = 3.90, $P_{Bonferroni} = 3.87 \times 10^{-4}$), but without the bias terms the decision preference favored classifying the trials as blue-dominant (DP = -0.21 ± 0.01; t(212) = -22.43, $P_{Bonferroni} < 0.001$), indicating classification based on sensory evidence alone provided conflicting information (**Fig. 3C**).

Applying the same analysis to blue-dominant trials, in Subtype 1, the strength of the sensory evidence alone was not sufficient to correctly classify the trials (DP = 0.09 ± 0.01; t(38) = 7.58, $P_{Bonferroni} = 1.23 \times 10^{-8}$) and this was compounded when including the bias terms (DP = 0.34 ± 0.01; t(38) = 32.05, $P_{Bonferroni} < 0.001$; **Fig. 3D**). Similar findings were observed in Subtype 2 (**Fig. 3E**). Moreover, in Subtype 3 inclusion of the bias terms resulted in a subset of trials favoring a red classification (DP = -0.13 ± 0.01; t(441) = -13.66, $P_{Bonferroni} < 0.001$) compared to when the classification decision was based only on the strength of sensory evidence (DP = -0.36 ± 0.01; t(441) = 45.00, $P_{Bonferroni} < 0.001$; **Fig. 3F**).

Additionally, we combined all the trials together to examine where the trials in each subtype reside along the evidence boundary. As can be observed in Figure 4A, subtypes were separated by the evidence boundary, linking trial-level variability to classification accuracy (**Fig. 4B**). Together, the results suggest that the bias term is a critical factor in decision-making particularly when sensory evidence is weak or provides conflicting information.

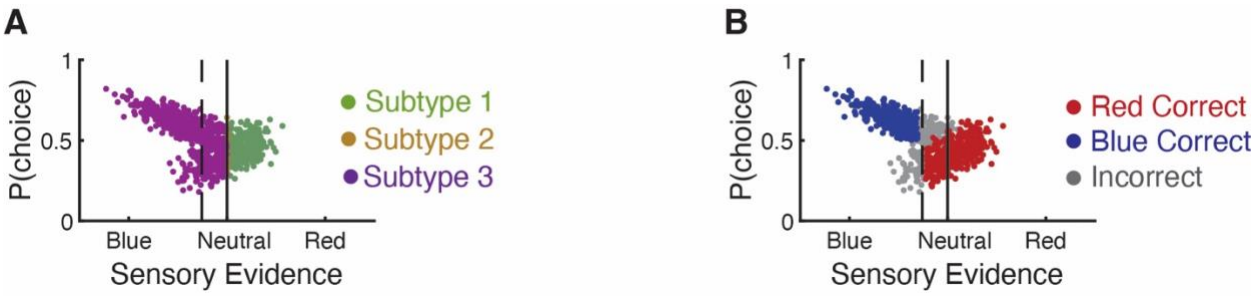

Figure 4. Evidence boundary separates subtypes. (A) Trial subtypes reside along distinct parts of the sensory evidence spectrum. Individual dots correspond to the trials. Dark black line corresponds to the evidence boundary. Dashed line corresponds the shift reflecting the bias term, C_{blue}. (B) Same as panel A but colored based on the accuracy of the classification. Red and blue dots reflect trials that were correctly classified as red or blue-dominant, respectively. Gray dots reflect trials that were incorrectly classified.

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

Additionally, noise is a challenge that both biological and artificial systems must contend with. We therefore tested how the bias terms affect classification performance as internal noise increases. Specifically, for each trial in the test set we added Gaussian noise, $\mathcal{N}(0, \sigma)$, either to the FC-2 activations or to the weights associated with mapping internal representations to outputs. We then quantified the classification accuracy with and without the bias terms. We observed that noise in the weights was more detrimental to classification accuracy than noise in the activations (**Fig. S2**). Moreover, the bias term preserved classification performance compared to when the bias term was not included when noise was added to the weights. On the other hand, when noise was added to the activations, the bias term had minimal impact on classification accuracy. The noise analyses suggests that the observed decision modes are not descriptive artifacts but have causal consequences for performance, with bias stabilizing decisions when the evidence readout itself is unreliable.

<u>Corroborating the findings in additional CNN models</u>

Lastly, we corroborated these findings in four additional CNN models. As can be observed in **Fig. S3-S6**, in all models, trial-level variation in activity could be clustered into distinct subtypes. Moreover, the bias term was a critical component in decision-making. Collectively, these results reveal that trial-level variation in activity reflects evidence-, uncertainty-, and bias-driven decision-making.

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). However, how such fluctuations relate to decision-making and behavior is not well understood. Here, we address this gap in knowledge using a convolutional neural network (CNN) trained to perform a perceptual decision-making task. A central contribution of this work is that it provides a mechanistic interpretation of multiple neural activation patterns observed during perceptual decision-making. Leveraging an established analytical framework, we found that a CNN exhibited fluctuations in activity that could be clustered into three discrete states mirroring multiple activation patterns observed in human brain activity during the same perceptual decision-making task. Our results demonstrate how such discrete activation states can arise naturally from continuous sensory evidence and correspond to distinct decision modes. Specifically, we show that different activation patterns reflect decisions dominated by strong sensory evidence, decisions made under low discriminability, and decisions in which bias plays a dominant role. This framework provides an interpretation of what distinct neural activation patterns may represent computationally and offers a principled explanation of how markedly different patterns of brain activity can coexist with similar behavioral performance.

The discrete states observed in the CNN suggest that the multiple patterns of activation seen in humans performing the same perceptual decision-making task may similarly reflect decisions made under different levels of sensory evidence. In particular, these results provide insight pertaining to the role played by the Default Mode Network (DMN) in decision-making. In the brain, increased engagement of the DMN is associated with states in which external inputs contribute little to brain activity (19). Analogously, in the CNN, Subtype 2 corresponds to a state in which the stimulus contributes little to activity. Nevertheless, decisions can still be generated both during DMN-dominant periods in humans and in Subtype 2 of the CNN. We therefore speculate that DMN engagement may be associated with decisions that rely more on choice-bias than on sensory evidence, paralleling the bias-driven decisions observed in the CNN.

Trial-to-trial variability in brain activity can be driven by many factors such as top-down processes (7), recurrent connectivity in local circuits (8), or bottom-up processes (9). In CNNs, by contrast, trial-to-trial variability is generated exclusively by bottom-up, stimulus-driven computations. Our results indicate that even in a noiseless system such as a CNN, the stimulus can generate distinct activation patterns, including a striking state characterized by the absence of any activity. This suggests that stimulus-driven dynamics may be a primary contributor to the emergence of multiple patterns of activation in perceptual decision-making (4). This is notable because fluctuations in neural activity during perceptual decision-making are often attributed primarily to factors such as top-down modulation rather than to the stimulus itself (20).

Modeling using signal detection theory suggests that choice-bias reflects a shift in the criterion (21). In humans, choice-bias has been shown to play a role in decision-making (22) and is thought to arise from experimental factors and measures are taken to minimize the effect (23). However, choice-bias can be beneficial in cases where one alternative is preferred over another (24). In humans, choice-bias may play a similar role as in the CNN and may not simply be a confounding factor (25, 26). How choice-bias 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 (27).

Beyond perceptual decision-making, these findings have practical implications for artificial neural networks (ANN) in general. In particular, the bias term is not merely an offset; it can meaningfully shape decisions, consistent with prior work showing that the bias term can make substantial contributions to prediction accuracy. Specifically, using only the bias term can achieve roughly 30–40% of full-model performance and backpropagating only the bias is sufficient to reconstruct an image (28). Consequently, when evidence is weak, ANNs may still produce confident-looking predictions that are bias-driven. This is clearest in Subtype 2, where removing the bias term causes the decision preference to collapse to neutral, making the bias critical for classification. More generally, comparing evidence-only logits (weights × activity) to full logits (weights × activity + bias) can act as a diagnostic to stratify errors. Additionally, other forms of bias may be encoded in the activity patterns of neurons as the input is transformed through successive layers of a CNN or ANN in general. These findings motivate explicit uncertainty handling and determining how the bias term affect outputs.

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 activation patterns (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, 100,000 unique stimuli were generated for training, and an additional 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 100,000 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 tests. All training and testing were conducted using Google Colab on an NVIDIA A100 GPU.

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 (29). 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 (30). We have previously used modularity maximization to identify multiple activation patterns in brain activity from fMRI data (4) and from EEG data (31, 32). 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 (33). In our previous work of clustering single trials, we had found that 100 iterations were sufficient to identify stable clusters (34). 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(35).

<u>The interplay between sensory evidence and choice-bias in the classifier</u>

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):

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

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 $\operatorname{argmax}\{P(\text{red}), P(\text{blue})\}$. 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. We used DP to dissociate the contribution of stimulus-driven evidence and choice-bias by recomputing the logits after setting the bias terms to zero ($C_{\text{red}} = C_{\text{blue}} = 0$). 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.

Code and data availability

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 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/).

Acknowledgments

We would like to thank Dobromir Rahnev for his thoughtful comments and suggestions. This research was supported by the U.S. Army DEVCOM Army Research Laboratory through mission funding (JOG) and army educational outreach program (JN, W911SR-15-2-0001). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army DEVCOM Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

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. 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.

20. C. D. Gilbert, W. Li, Top-down influences on visual processing. *Nature reviews neuroscience* **14**, 350–363 (2013).

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

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

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

24. 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).

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

26. 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).

27. 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).

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

29. 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).

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

31. 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).

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

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

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

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