A layer-resolved diagnostic identifies bias-driven decisions in deep neural networks

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: Deep Neural Networks, Interpretability, Attribution Methods, Decision-Making, Bias

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

Abstract

Deep neural networks can appear confident even when weakly supported by inputs. Current approaches such as calibration and feature attribution relate confidence to error or map predictions to features, but do not quantify how the decision margin is supported. Here we introduce a layer-resolved Bias Dominance Index (BDI) that decomposes the margin into feature and bias components to localize where bias-dominance emerges. Across models – convolutional neural networks (CNN) and vision transformers (ViT)– BDI increases as feature support weakens and high confidence can coexist with bias-dominant decisions. Moreover, bias-dominance emerges in convolutional layers in CNNs and normalization layers in ViT. Perturbation analyses further show that bias components can be beneficial when readout weights are degraded. Finally, we operationalize decision composition as a two-parameter acceptance rule that combines confidence with BDI for auditing. Together, these results position BDI as a diagnostic for monitoring and triage, distinguishing feature- versus bias-driven decisions.

Introduction

Deep neural networks (DNNs) increasingly operate as decision-making systems in settings where inputs are incomplete, ambiguous, or degraded – images are occluded (1), sensor streams are noisy (2, 3), and language inputs are underspecified (4, 5) or adversarially framed (3, 6). In these situations, models often remain decisive, producing high-confidence predictions even when the input provides limited support (7). For deployment, the central question is not only whether a prediction is confident, but what computational substrate supports that confidence. Specifically, is the decision driven by an input-dependent feature component, or does it primarily reflect input-independent component structure internal to the model, such as offsets (i.e. bias) and learned baselines that function as implicit priors (8).

A substantial body of work provides complementary tools for probing these issues. Uncertainty estimation and calibration relate confidence to error (9, 10), and attribution methods map outputs to influential input features or internal units (11–14). Building on these foundations, we focus on a different but related quantity, the composition of the decision margin itself. In particular, we ask how much of the margin is supported by feature-driven computations versus the bias component. This compositional perspective matters because confidence can, in principle, be supported primarily by the bias component, yielding decisions that appear decisive despite weak feature support, a scenario that is not apparent from confidence alone. Here, bias denotes an intercept-like contribution to the logit margin that is independent of the current input. Importantly, bias is an input-independent offset, rather than bias arising from training-set composition, label imbalance, spurious correlations, systematic miscalibration or dataset artifacts.

To accomplish this, we introduce a framework that interrogates the decision margin itself. For a broad class of neural networks, the decision margin can be expressed as an affine function of internal representations, which admits an explicit decomposition into a feature component and bias component (**Fig. 1A–C**). We leverage this structure to define a Bias Dominance Index (BDI), quantifies the relative support of the margin by input-independent offsets versus input-linked computations (**Fig. 1D,E**). BDI is not a calibration metric or a default predictor of accuracy, but a mechanistic diagnostic that distinguishes decisions whose margins are primarily

feature-supported from those in which feature support is weakened and offsets disproportionately shape the output. Crucially, our goal is not to restate that bias terms exist, but to use their *relative dominance* to identify decisions that are primarily feature-driven versus those in which feature-driven contributions are attenuated, and bias disproportionately shapes the output (**Fig. 1F**).

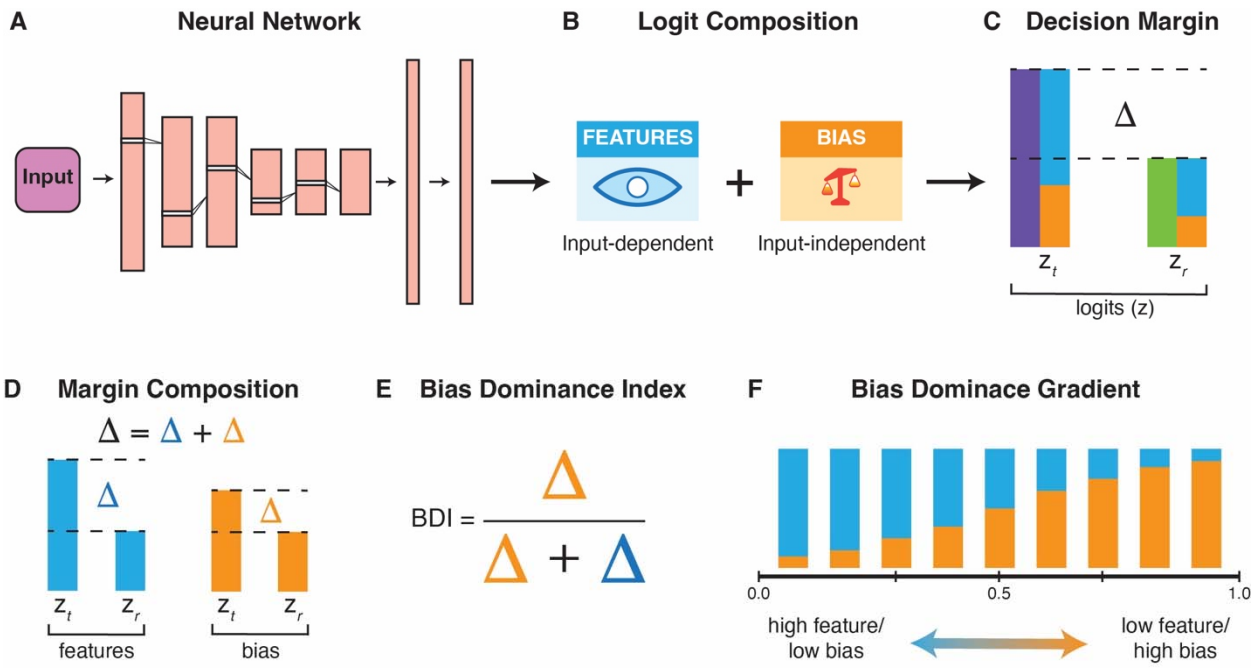

Figure 1. Decision-margin decomposition and bias dominance index. (A) Schematic of a feed-forward neural network that maps an input to an output through successive processing layers. (B) Conceptual separation of the final logit signal into input-dependent feature and input-independent bias components. (C) The decision margin is defined by the logit difference (Δ) between the top predicted class *t* (purple) and its closest competitor *r* (green); blue and orange bars denote the feature and bias contributions to the logits for *t* and *r*, respectively. (D) The decision margin is partitioned into feature and bias contributions for the predicted class *t* and closest competitor *r*. (E) The Bias Dominance Index (BDI) quantifies how much of the margin is supported by bias versus features. (F) Illustration of how BDI varies across inputs/decisions. Low BDI corresponds to feature-grounded margins, whereas high BDI indicates margins that are increasingly bias-supported.

Furthermore, we extend the analysis to derive a layer-resolved BDI by attributing each layer's contribution to the decision margin and decomposing that contribution into feature and bias components. This yields a depth-dependent map of the decision margin, enabling localization of bias-dominance to specific layers of the network. Additionally, we adapt the framework to multi-

class models using a label-free decision margin defined by the top-1 versus top-2 logit difference, because deployment settings often lack labels enabling BDI-based auditing of decisions without ground-truth annotations.

We apply these diagnostics to a set of convolutional neural networks (CNNs) trained on a controlled perceptual decision-making task. BDI reveals systematic decisions in which feature contributions are attenuated and the bias component disproportionately supports the margin, while layer-resolved BDI localizes where this mode emerges across network depth. Additionally, we test mechanistic relevance of the bias component through perturbations that selectively degrade feature-carrying computation. Finally, we establish that the BDI signature generalizes to other convolutional (AlexNet (15) and VGG-16 (16)) and vision transformer (ViT-b/16 (17)) models. Similarly in these networks, BDI flags bias-dominant yet confident decisions, localizes where bias dominance arises in the network hierarchy, and provides a framework for auditing and stress-testing models under degraded feature and distributional uncertainty. Lastly, we operationalize decision composition into a two-parameter acceptance rule that combines confidence with BDI for auditing. Together, these results position BDI as a practical diagnostic for interpretability and reliability in modern neural networks.

Results

<u>Decision decomposition in a controlled dot-discrimination task</u>

To test how internal computations support the decision margin, we trained five CNNs on a controlled dot-discrimination task in which each image contained intermixed red and blue dots and the model reported the dominant color.

Across models, performance was above chance (Accuracy = 0.75 ± 0.0038; $t(4) = 65.41$, $P = 3.27 \times 10^{\square\square}$) and comparable to human accuracy on the same task (Accuracy = 0.77 ± 0.0095; $t = 1.84$, $P = 0.07$; **Fig. S1**). Accuracy was above chance for both red-dominant (Accuracy = 0.75 ± 0.035; $t(4) = 7.91$, $P = 1.38 \times 10^{\square 3}$) and blue-dominant trials (Accuracy = 0.72 ± 0.034; $t(4) = 5.65$, $P = 0.004$), with no preference for either color ($t(4) = 0.70$, $P = 0.52$).

We then quantified how each model's decision margin was constructed from the feature component versus the bias component by computing the Bias Dominance Index (BDI) at the last affine layer. Across all five models, the output-layer BDI ranged from weakly to strong driven by bias (BDI_{IQR} = 0.291 ± 0.036), revealing marked trial-to-trial heterogeneity in how the decision margin was supported (model-specific values are shown in **Fig. S2**). Moreover, on average, 35.48 ± 8.71% of red-dominant trials and 47.80 ± 14.05% of blue-dominant trials exceeded the bias-dominance threshold (BDI > 0.5), indicating that the bias component frequently contributed more to the margin than the feature component. Notably, even at similar confidence levels, trials exhibited widely different BDI values, with some high-confidence responses remaining bias-dominant, demonstrating that decisiveness can coexist with a small contribution from the feature component to the decision margin (**Fig. 2**). Together, these results suggest that the bias component can substantially shape decision outcomes.

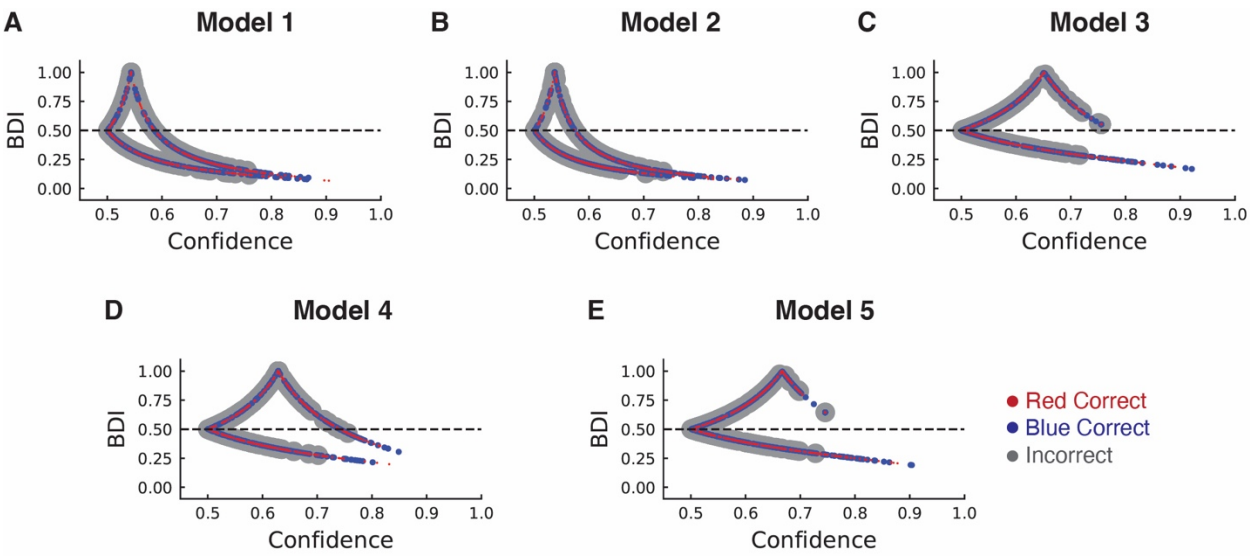

Figure 2. Bias can produce confident-looking decisions at the output layer. (A) Relationship between BDI and confidence in Model 1. BDI was computed from the red–blue logit margin at the output affine layer. Confidence is estimated as the maximum *softmax* probability (i.e., the probability assigned to the chosen class) and is neutral at 0.5 for a two-choice perceptual decision task. Each dot corresponds to a trial, colored by outcome. The horizontal dashed line marks the bias-dominance criterion (BDI = 0.5). (B-E) Same as panel A, but for Model 2-5, respectively.

Layer-resolved bias dominance index on the dot-discrimination task

The bias component is not confined to the output layer; its influence can propagate through the processing hierarchy and shape the decision margin. To quantify how bias dominance evolves with depth, we computed layer-resolved BDI across all convolutional and fully connected layers in the dot-discrimination task, yielding a trial-resolved depth profile of bias dominance. This analysis treats each layer as a potential locus where the bias component can modulate intermediate representations, enabling us to determine how bias and feature components shape the decision margin.

Across models, layer-resolved BDI showed structured depth dependence (**Fig. 3A–E**). The fraction of bias-dominant trials (BDI > 0.5) was highest in early convolutional layers ($BDI_{Feature\text{-}0}$ = 47.10 ± 1.82%; t(4) = 24.87, P_{FDR} = 1.05 × 10□□) and re-emerged in late linear layers ($BDI_{Linear\text{-}4}$ = 33.14 ± 8.20%; t(4) = 4.04, P_{FDR} = 0.04; $BDI_{Linear\text{-}6}$ = 41.66 ± 8.15%; t(4) = 5.11, P_{FDR} = 0.02; **Fig. 3F**), indicating that bias-related support can arise upstream and persist to the decision margin.

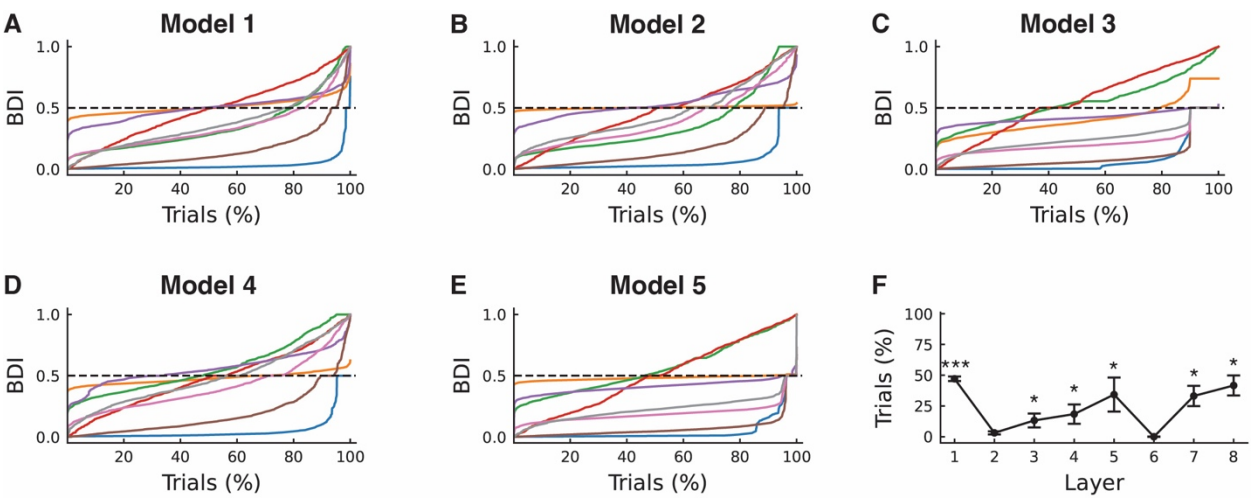

Figure 3. Layer-resolved Bias Dominance Index across network depth. (A) Layer-resolved BDI profile for Model 1. For each layer, trials are ranked by BDI values. Colored lines correspond to layers spanning convolutional feature and fully connected linear stages that contain a bias component. For visualization purposes we have omitted the layer labels, but layer specific names are provided in Table S1. The dashed line indicates the bias-dominance criterion BDI = 0.5. (B–E) Same as panel A for Models 2-5. (F) Percentage of trials with BDI > 0.5 across layers. Plot shows Mean ± SEM across models. Layers are ordered sequentially by execution,

and only layers with bias terms are included. Statistical significance was estimated using one-sample *t*-test and FDR-corrected. ***$P_{FDR} < 0.001$ and *$P_{FDR} < 0.05$.

<u>Retaining the bias term preserves performance against noise in the weights</u>

Real-world decision systems must operate under noise and corrupted internal signals. We therefore asked whether the bias component contributes causally to performance when the feature component is degraded. For each test trial, we added Gaussian noise, $\mathcal{N}(0, \sigma)$, either to the neural activation vector or to the classifier weights mapping the representations to output logits, and quantified accuracy with and without the bias component ($b_{red} = b_{blue} = 0$). Using both perturbations distinguishes sensitivity to representational noise (activation corruption) from sensitivity to readout uncertainty (weight corruption), isolating whether bias support stabilizes decisions under internal degradation of different kinds.

In a representative model (Model 4), weight noise impaired accuracy more than activation noise ($P_{FDR} < 0.001$; **Fig. 4**). Under weight noise, retaining the bias component improved performance relative to removing it (Accuracy = 0.55 ± 0.008 vs 0.49 ± 0.006; t(9) = 28.56, $P_{FDR} = 3.84 \times 10^{-1}$). Under activation noise, accuracy did not differ with versus without the bias component (Accuracy = 0.59 ± 0.017 vs 0.59 ± 0.015; t(9) = 1.94, $P_{FDR} = 0.24$). Similar results were observed in other models (**Fig. S2**), indicating that the bias component can stabilize decisions when readout weights are unreliable.

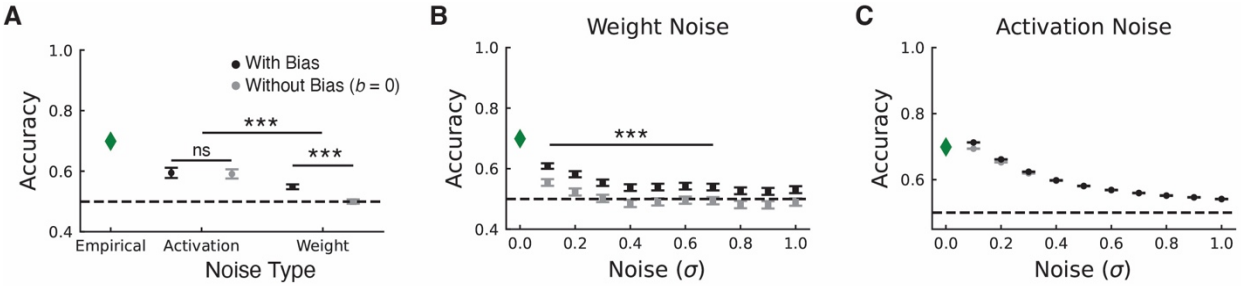

Figure 4. Bias component preserves performance against noise in the weights. (A) Average classification accuracy after adding noise. Green diamond represents the empirical performance with no added noise ($\sigma = 0$) is shown for reference. Black and gray dots represent the average classification value across all noise levels for with and without bias term ($b_{red} = b_{blue} = 0$), respectively. The dashed line indicates chance performance (0.5). Error bars show Mean ± SEM. (B) Classification accuracy for each level of noise added to the weights. (C)

Same as panel B, but noise added to the neuronal activations. Statistical significance was determined using paired-sample *t*-test and FDR corrected. ***$P_{FDR} < 0.001$; ns = not significant.

<u>Generalization to ImageNet-pretrained networks using a label-free decision margin</u>

To assess whether bias dominance extends beyond the controlled perceptual decision-making task, we evaluated ImageNet-pretrained AlexNet, VGG-16, and ViT-b/16 on Tiny ImageNet and Tiny ImageNet-C images. These models produce 1000-way ImageNet logits. We therefore used this to determine if BDI can be utilized in a label-free manner. Specifically, we quantified the decision support using a label-free margin defined as the difference between the highest and second-highest logits, $\Delta z = z_{(1)} - z_{(2)}$, which does not require class labels. Competitor indices $(1), (2)$ were determined per input by the model's own ranking.

Ground-truth labels were not used to define Δz or compute BDI, but labels were used only for post-hoc visualization and mismatch-rate analysis. Label match / mismatch on Tiny ImageNet were computed only after mapping Tiny ImageNet classes (wnid) into the ImageNet-1K (synset) space. Specifically, each Tiny ImageNet wnid is mapped to its corresponding ImageNet-1K class. We first focus on the relationship between BDI and confidence in output layer since in this layer the relationship between the feature component, C_{feat}, and the bias component directly influences the decision margin.

In AlexNet, BDI varied widely across inputs (Mean = 0.114; Std = 0.161; IQR = 0.111; **Fig. S4**). In total, 4.47% of trials exceeded the bias-dominance threshold (BDI > 0.50). To summarize the relationship between decision composition and confidence, we overlaid confidence-binned medians and interquartile ranges of BDI on the trial-level scatter, which revealed a strong inverse trend but substantial within-bin variability at low-to-moderate confidence (**Fig. 5A**).

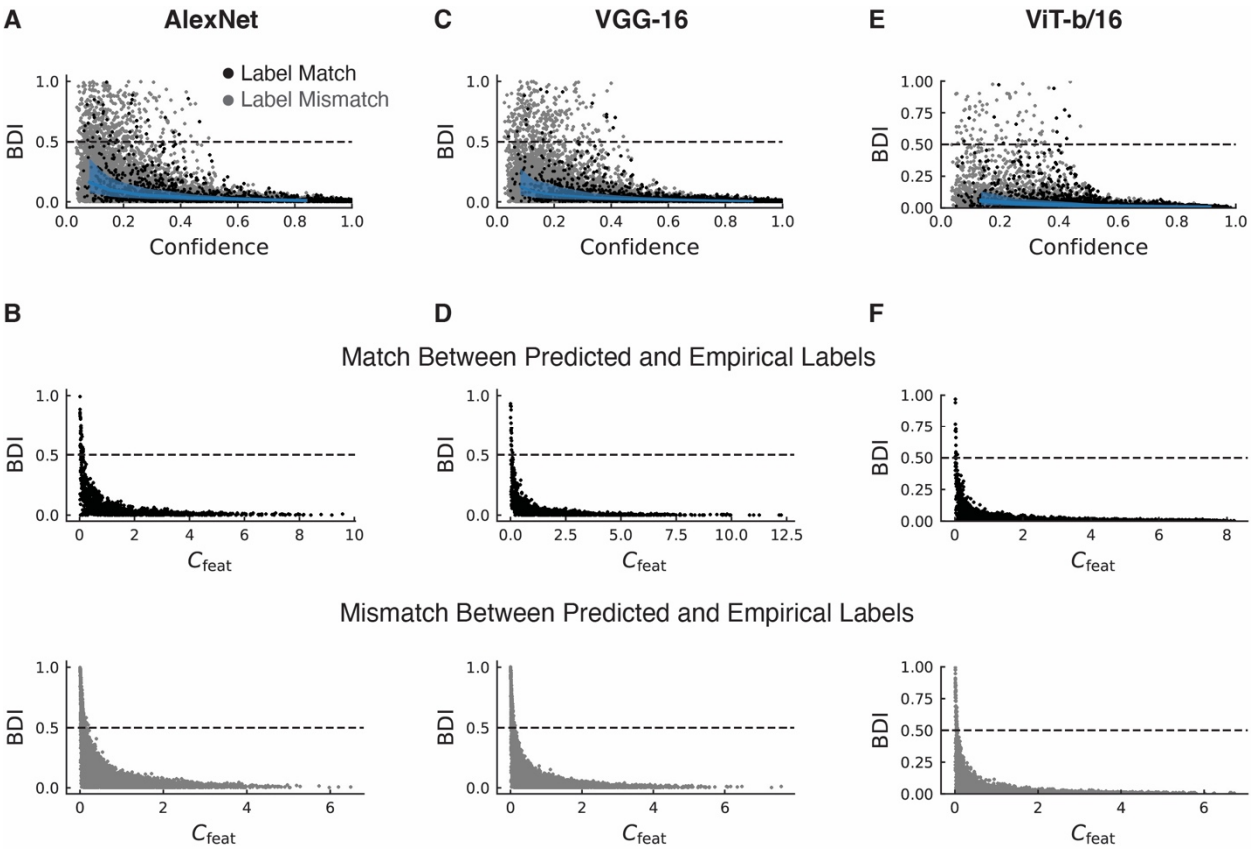

Figure 5. Bias-dominant decision margins persist at high confidence. (A) Relationship between BDI and confidence for AlexNet. Solid line shows the median BDI within confidence bins (N_{bins} = 10) and the shaded band indicates the interquartile range, summarizing the central trend and within-confidence variability in decision composition. BDI was computed from the top-1 versus top-2 logit margin at the affine output layer. For visualization, points are colored by label match (black) and mismatch (gray). Confidence is the softmax probability assigned to the chosen class (chance baseline 1/1000 = 0.001 for 1000-way ImageNet outputs). Analysis is based on 10,000 images from the Tiny ImageNet dataset. Label match/mismatch is synset-mapped; Tiny ImageNet wnid → ImageNet□1K; used for reporting only. The horizontal dashed line marks the bias-dominance criterion BDI = 0.5. (B) Relationship between BDI and the feature component (C_{feat}). (C,D) Same as panel A and B, but for VGG-16. (E,F) Same as panel A, but for ViT-b/16.

Importantly, the confidence–BDI relationship alone does not exclude cases in which the contribution of the feature component, C_{feat}, to the decision margin is large in absolute terms, yet the bias component is larger still. In such cases, BDI remains high despite substantial feature component. We therefore related BDI to C_{feat}, to test whether high BDI primarily reflects reduced feature support for the decision margin, rather than simply an increase in bias magnitude.

Relating BDI to the feature component showed that BDI increased as feature support decreased (**Fig. 5B**). Even in high-BDI/low-C_{feat} cases, top-1 predictions could still agree with the remapped label, indicating that label agreement can occur despite weak feature support. VGG-16 and ViT-b/16 showed the same qualitative pattern (**Fig. 5C–F; Fig. S4**).

Layer-resolved analyses under the label-free margin localized where bias-dominant support emerges across architectures. In AlexNet and VGG-16, bias dominance was most pronounced in convolutional layers (**Fig. 6A–D**), whereas in ViT-b/16 it was strongest in LayerNorm modules (**Fig. 6E,F; Fig. S5**). Together, these results indicate that bias dominance increases as feature support weakens, can still coincide with synset-mapped label agreement, and bias dominance is a depth-distributed computational property across architectures, positioning BDI as a label-free diagnostic for identifying and localizing decisions that are disproportionately bias-supported rather than feature-supported.

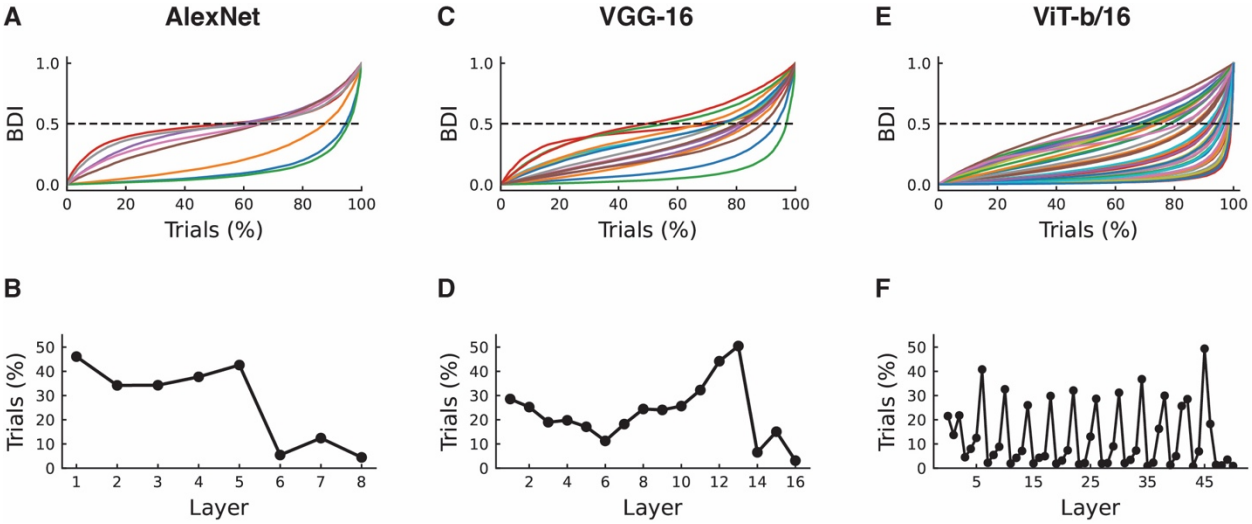

Figure 6. Layer-resolved bias dominance localizes margin support. (A) Layer-resolved BDI profile for AlexNet. Colored lines correspond to layers spanning convolutional and fully connected layers that contain a bias component. Analysis is based on 10,000 images from the Tiny ImageNet dataset. Trials are ranked by BDI values. The dashed line indicates the bias-dominance criterion BDI = 0.5. (B) Percentage of bias-dominant trials with BDI > 0.5 across layers for AlexNet. Layers are ordered sequentially according to their execution order on the input, and only layers with bias components are included. (C) VGG-16 results shown as in panel A. (D) Percentage of bias-dominant trials with BDI > 0.5 across layers for VGG-16. (E) ViT-b/16 results shown as in panel A. (F) Percentage of bias-dominant trials with BDI > 0.5 across

layers for ViT-b/16. For visualization purposes we have omitted the layer labels, but layer specific names are provided in Table S1.

Having established that BDI captures depth-distributed shifts from feature- to bias-supported margins even in the presence of synset-mapped label agreement, we next tested whether this signature becomes more prevalent under explicit distribution shift induced by corruptions. Under Tiny ImageNet-C corruption, AlexNet again exhibited broad BDI values and increased prevalence of bias-dominant trials (6.80% with BDI > 0.50; **Fig. 7A**). When we matched trials by *softmax* confidence, mismatch rates were comparable across low- and high-BDI strata across confidence bins (t(19) = 0.50; P = 0.62; N = 20 confidence bins; **Fig. 7B**). This does not imply that BDI is redundant; rather, it shows that confidence and accuracy-like outcomes do not identify *how* the model produced a decisive margin. At the same confidence level, the same predicted label can be supported primarily by features or by bias, indicating that mismatch-rate and confidence alone are insufficient to determine the dominant decision mechanism on a given trial. The same patterns was observed for VGG-16 (**Fig. 7C,D**) and ViT-b/16 (**Fig. 7E,F**).

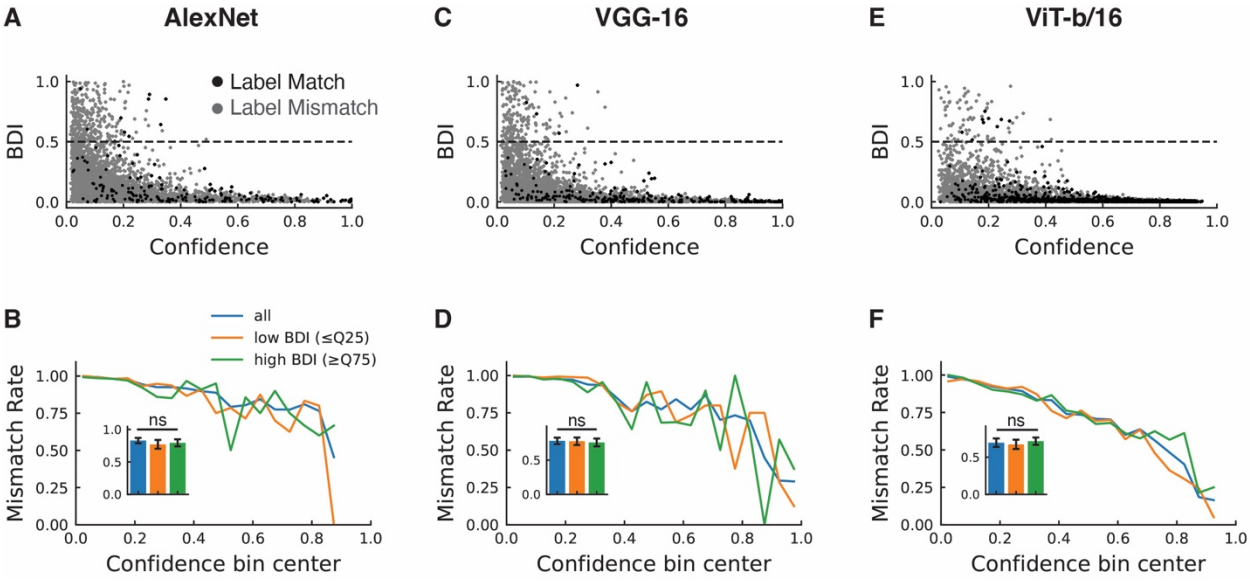

Figure 7. Bias-dominant margins persist under input corruption. (A) Relationship between BDI and confidence for AlexNet on Tiny ImageNet-C. BDI was computed from the top-1 versus top-2 logit margin at the affine output layer. Points are colored by synset-mapped label match (black) versus mismatch (gray) for reporting only (Tiny ImageNet wnid → ImageNet-1K).

Confidence is the *softmax* probability assigned to the chosen class (chance baseline 1/1000 = 0.001 for 1000-way ImageNet outputs). Horizontal dashed line denotes the bias-dominance criterion (BDI = 0.5). (B) Label mismatch rate as a function of confidence. Trials are binned by confidence and within each confidence bin trials are partitioned by BDI (bottom vs top quartile), yielding mismatch-rate curves for low-BDI and high-BDI subsets alongside the aggregate ("all") curve. Inset shows the Mean ± SEM across bins. (C,D) Same as panels A and B, but for VGG-16. (E,F) Same as panels A and B for ViT-b/16. Statistical significance was estimated using paired-sample *t*-test. ns = not significant.

<u>Confidence–BDI decision policy for accepting predictions</u>

To translate decision composition into an operational audit rule, we evaluated a two-parameter, label-free acceptance policy that retains a prediction only when it is both decisive by standard uncertainty criteria and not bias-dominant by the decision-composition diagnostic. This motivates using BDI not only as a descriptive diagnostic, but as a label-free triage axis that targets decision mechanism rather than confidence alone. Specifically, we accepted trials satisfying Confidence $\geq \tau$ and $BDI_{all} \leq \gamma$, where τ sets a minimum confidence threshold and γ limits tolerance for bias-supported margins. In AlexNet, sweeping (τ, γ) yields the fraction of predictions accepted (**Fig. 8A**).

Importantly, this policy view is mechanism-aware and fills a critical gap because tightening γ preferentially excludes decisions whose margins are primarily bias-supported even at high confidence. This is illustrated with a representative high-confidence trial whose predicted label matched the empirical labels, yet attribution analysis using DeepLIFT did not exhibit a clear correspondence between salient features and either the decision or its confidence (13) (**Fig. 8B**). Attribution methods can highlight regions that influence the output, but they do not directly quantify how the model's decisiveness is supported, leaving ambiguous how high confidence is warranted when the highlighted features appear weak or misaligned. By contrast, BDI indicated that these margins were primarily bias-supported, with C_{bias} dominating C_{feat}. These observations were replicated in both VGG-16 (**Fig. 8C,D**) and ViT-b/16 (**Fig. 8E,F**). Thus, rather than assuming that bias-dominant decisions are always incorrect, the Confidence–BDI policy provides a label-free axis for triage and auditing that can flag confident-but-bias-dominant predictions for additional scrutiny.

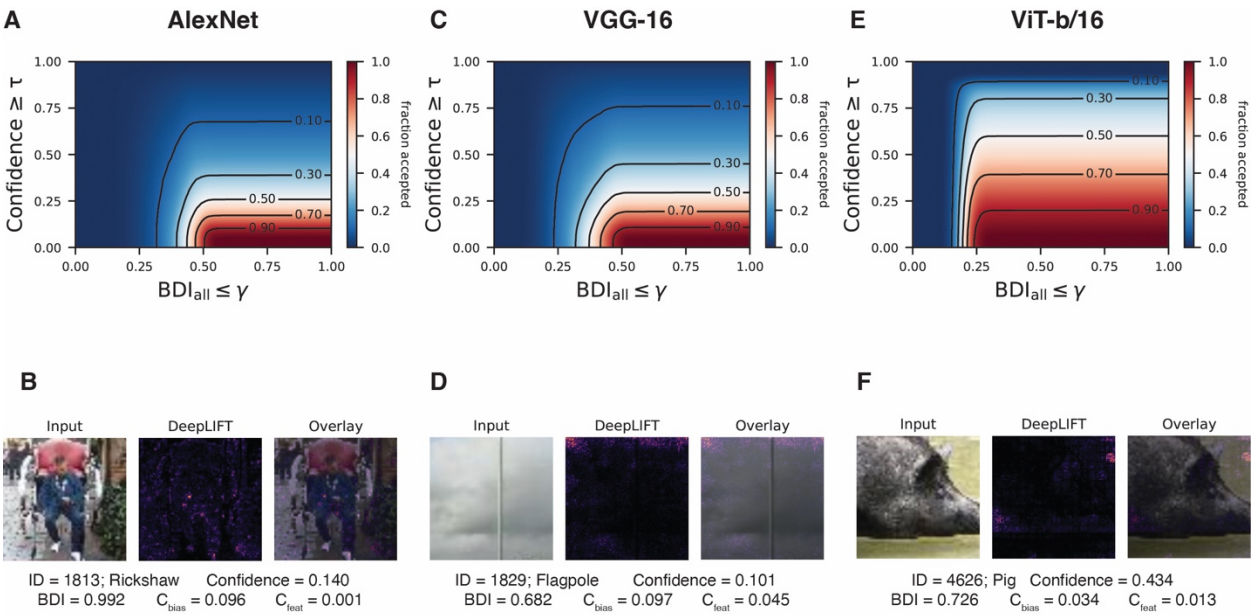

Figure 8. Confidence–BDI acceptance policy for mechanism-aware triage. (A) Coverage heatmaps over the (τ, γ) grid for AlexNet where trials are accepted if Confidence ≥ τ and $BDI_{all} \leq \gamma$. Contours indicate iso-coverage levels. (B) Illustrative trials with high-confidence and high-BDI value from the output layer whose top-1 predictions mapped to the empirical labels (synset-mapped). Despite mapped label match, the decision margin is weakly supported by the feature component (low C_{feat} relative to C_{bias}), DeepLIFT (13) attribution corroborates the limited influence of the feature component on the decision margin. DeepLIFT identifies influential regions, but it remains ambiguous how those regions support the model's high confidence. Confidence is the *softmax* probability assigned to the chosen class (chance baseline 1/1000 = 0.001 for 1000-way ImageNet outputs). (C,D) Same as panel A and B, but for VGG-16. (E,F) Same as panel A and B, but for ViT-b/16.

<u>Relationship between BDI and bias magnitude</u>

A potential concern is that BDI could be viewed as a re-expression of the bias contribution, since when feature support is attenuated ($C_{\text{feat}} \approx 0$) the index necessarily approaches its bias-dominant limit. However, BDI is not intended to quantify the absolute magnitude of bias, but rather the relative support of the decision margin by bias versus features. To test whether BDI collapses to a bias-only statistic, we computed the bias contribution to the top-1 versus top-2 margin at each layer, $C_{\text{bias}}^{(\ell)}$, and examined its association with layer-resolved BDI. The relationship between BDI and $|\, C_{\text{bias}}^{(\ell)} \,|$, exhibited substantial dispersion in BDI values at fixed $|\, C_{\text{bias}}^{(\ell)} \,|$ (**Fig. S6–S8**). This pattern indicates that bias dominance is not determined by bias magnitude alone, but depends jointly on bias and feature support, consistent with BDI's ratio form.

Discussion

Deep neural networks (DNNs) can exhibit high confidence even when their decisions are only weakly supported by the underlying signal. Here we introduce the Bias Dominance Index (BDI), to provide a quantitative lens to see through the "confidence mask" of DNNs. BDI is derived by decomposing the decision margin into its constituent parts, an input-dependent feature component and input-independent bias. Across large-scale vision models – AlexNet, VGG-16, and ViT-b/16 – we find that bias dominance is widespread and layer-resolved. This pattern persisted even for high-confidence predictions, indicating that BDI provides an independent axis for mechanism-aware auditing and triage. In turn, BDI can be incorporated into acceptance policies that trade coverage against tolerance for offset-dominant margins beyond confidence alone.

These findings have practical implications for interpreting and auditing modern neural networks. Our results indicate that bias-like support can dominate across layers and networks may produce confident-looking predictions even when input-dependent support is not the dominant contributor to the decision margin. 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 (18).

The influence of bias is not confined to the readout layer but can emerge in early processing and persist through intermediate representations, highlighting that bias is a distributed property of computation. In convolutional neural networks, the influence of bias was strongest in the early layers, indicating that internal priors shape the feature extraction process from its inception. In biological perception, systematic biases are often interpreted as adaptive priors – shaped by natural stimulus statistics and evolutionary pressures–that help stabilize feature inference under uncertainty (19). In convolutional neural networks, however, these early offsets may represent architectural shortcuts that allow the model to bypass complex feature synthesis when a prior is sufficiently strong (20, 21). Moreover, in the transformer architecture, bias was introduced

primarily through layer normalization, pointing to normalization as a key locus where input-independent offsets can be amplified across depth. More broadly, because modern architectures include multiple normalization and other affine operations, the set of loci where input-independent offsets can enter or be amplified expands, underscoring the need for diagnostic measures. Overall, if one aims to reduce bias-dominant decision margin support, effective interventions may need to target early representational dynamics and normalization mechanisms—not only the output-layer decision rule.

While bias in machine learning is often framed as an issue to be corrected (22–24), our perturbation analyses reveal a more nuanced role. Under conditions of high noise or weight degradation, bias acted as support, preserving classification accuracy when evidentiary weights were compromised. This suggests that the bias in neural architectures may function as a learned internal prior, providing a baseline of robustness in volatile environments. The challenge for future research is to distinguish between *beneficial priors* that aid generalization and *harmful biases* that drive erroneous overconfidence.

Most deployment pipelines treat softmax confidence as a proxy for reliability and implement a one-dimensional acceptance rule, for example accept if confidence exceeds a threshold (25). Our results suggest that confidence alone can conflate qualitatively different factors underlying decisions. Predictions can be equally confident yet differ in whether the decision margin is supported by input-dependent feature or by input-independent bias. Instead, a two-parameter policy that jointly thresholds confidence and BDI could allow practitioners to trade coverage against tolerance for bias-dominant margins. Because both quantities are computed from internal decision variables, such a policy is label-free and could be deployed when ground truth is unavailable or delayed. More broadly, this perspective reframes “reliability” as not only a question of margin magnitude (confidence) but also of margin composition (feature vs bias), providing an interpretable handle for auditing and triaging predictions that would otherwise pass confidence screening.

Moreover, BDI can be estimated with label-free inputs (top-1 vs. top-2) expanding the practical utility of this framework. BDI could be applied in high-stakes deployments (e.g., medical

diagnostics (26)) where ground truth is unavailable or delayed. In such settings, BDI could serve as a monitor that flags high-confidence outputs whose margins are primarily offset-supported. This may provide a complementary audit signal for identifying predictions that warrant additional review under degraded or out-of-distribution inputs. This way, BDI could help flag cases at risk of overconfident errors, spurious predictions, over-extrapolation, and brittle generalization under degraded or out-of-distribution inputs.

Beyond machine learning, these findings may also inform computational neuroscience models of the visual cortex and decision-making. Our observation that bias-dominant margin support can emerge early and persist across depth may be a way to formalize how *priors* influence perceptual decisions when stimulus-linked support is weak. This framing may be relevant because CNNs have been widely used as mechanistic models for aspects of human vision, including object recognition (27, 28), representational structure (29) and processing in the visual cortex (30–32). Moreover, in prior work, humans performing the same dot-discrimination task exhibited multiple reproducible activation patterns, including engagement of the Default Mode Network (33). Default Mode Network engagement is associated with brain states in which internally generated dynamics dominate over externally driven (34). In this light, our results raise the possibility that these pattern shifts reflect changes in decision composition. Accordingly, we speculate that Default Mode Network engagement during decision-making may be preferentially associated with trials in which sensory features provide relatively little margin support and internal priors provide comparatively greater support, paralleling bias-dominant decision margins in DNNs.

By shifting the focus from input features to decision composition, we provide a foundation for auditing neural networks. Although we illustrate the approach in CNNs and vision transformers, the formulation is architecture-agnostic and can, in principle, be applied to any model whose decision variables can be expressed as affine functions of internal representations, including recurrent networks and large language models. Moreover, future work tracking the developmental trajectory of BDI over training could pinpoint when a model transitions from learning input-driven structure to exploiting bias. Taken together, rather than merely rationalizing outputs post hoc, the BDI framework exposes the balance between input-linked

support and model-intrinsic offsets, moving us toward AI systems whose decisions are not only accurate, but more readily interpretable.

Methods

<u>Decomposition of the decision margin and bias dominance index</u>

For each input x_i (trial i), logits are produced by an affine readout from an internal activation vector A_i from a total of K classes:

$$z_{i,k} = W_k^{\top} A_i + b_k, k \in 1, \dots, K.$$

We define the logit margin between a target class t and alternative class r as:

$$\Delta z_i^{(t,r)} = z_{i,t} - z_{i,r}.$$

The margin admits an exact decomposition into an input-dependent feature component ($\Delta z_{feat,i}^{(t,r)}$) and an input-independent bias component ($\Delta z_{feat,i}^{(t,r)}$):

$$\Delta z_i^{(t,r)} = \underbrace{(W_t - W_r)^{\mathrm{T}} A_i}_{\Delta z_{feat,i}^{(t,r)}} - \underbrace{(b_t - b_r)}_{\Delta z_{bias,i}^{(t,r)}}.$$

We define the Bias Dominance Index (BDI) for the decision margin between (t, r) on trial i as:

$$\mathrm{BDI}_i^{(\mathrm{t,r})} = \frac{|\ \Delta z_{bias,i}^{(t,r)}\ |}{|\ \Delta z_{bias,i}^{(t,r)}\ | + |\ \Delta z_{feat,i}^{(t,r)}\ | + \varepsilon}$$

with $\varepsilon > 0$ for numerical stability. By construction, $\mathrm{BDI} \in [0,1]$ values near 0 indicate feature-dominant margin, whereas values near 1 indicate margins dominated by bias. We use $\mathrm{BDI} > 0.5$

as a natural dominance threshold. Unlike margin magnitude or confidence alone, BDI isolates how a margin is constructed.

Layer-resolved BDI. To extend BDI beyond the output layer, an analogous feature–bias decomposition at multiple stages of the network is computed. This is accomplished by extending BDI across depth through attributing each layer's feature and bias contributions to the decision margin. Specifically, for each affine transformation at layer ℓ, the pre-nonlinearity output is:

$$u_i^{(\ell)} = W^{(\ell)} a_i^{(\ell-1)} + b^{(\ell)}.$$

We partition the output into a input-dependent feature component:

$$u_{feat,i}^{(\ell)} = W^{(\ell)} a_i^{(\ell-1)}$$

and a bias component (broadcast across spatial positions for convolutional layers):

$$u_{bias,i}^{(\ell)} = b^{(\ell)}.$$

In convolutional neural networks (CNN), the input-independent component, $u_{bias,i}^{(\ell)}$ corresponds to the additive bias parameters in affine layers. In the vision transformer network, ViT-b/16, there is an additional input-independent bias component (i.e. offset) in the LayerNorm module – β parameter. We treat the LayerNorm shift (β) as an input-independent offset because it contributes an additive term to the post-normalization representation that does not depend on the current input sample. In the label-free decision margin $\Delta z = z_{(1)} - z_{(2)}$, the contribution of β can be written as $\Delta z_{\mathrm{LN}} = (w_{(t)} - w_{(r)})^{\top}\beta$. Note that, the Δz_{LN} term is independent of the current input, is conditional on the trial's competitor classes (1) and (2). When computing label-free margin support (top-1 versus top-2), offsets are propagated through subsequent affine projections and can contribute to the decision margin even in the absence of strong input-linked support. Accordingly, we include β-derived contributions in the bias component for ViT (derivation in the Supplemental).

The gradient of the output margin with respect to the layer's pre-nonlinearity activations is defined as:

$$g_i^{(\ell,t,r)} = \frac{\partial \Delta z_i^{(t,r)}}{\partial u_i^{(\ell)}}.$$

The gradient-weighted feature and bias contributions at layer ℓ is estimated as:

$$C_{feat,i}^{(\ell,t,r)} = \langle g_i^{(\ell,t,r)}, u_{\text{feat},i}^{(\ell)} \rangle$$

$$C_{\text{bias},i}^{(\ell,t,r)} = \langle g_i^{(\ell,t,r)}, u_{\text{bias}}^{(\ell)} \rangle$$

where $\langle \cdot,\cdot \rangle$ denotes the inner product (includes all spatial positions for convolutional layers). The layer-resolved BDI is:

$$\text{BDI}_i^{(\ell;t,r)} = \frac{\mid C_{bias,i}^{(\ell,t,r)} \mid}{\mid C_{bias,i}^{(\ell,t,r)} \mid + \mid C_{feat,i}^{(\ell,t,r)} \mid + \varepsilon}.$$

Critically, the layer-resolved diagnostic identifies specific layers where the bias dominance increases most sharply and therefore provides a principled target for interventions.

Additionally, we define a summary statistic, BDI_{all}, that compresses the layer-wise decomposition into a single measure of overall decision composition. For input i, we pool contribution of bias and feature components across layers by summing $\mid C_{bias,i}^{(\ell;t,r)} \mid$ and $\mid C_{feat,i}^{(\ell;t,r)} \mid$ over $\ell = 1, \ldots, L$:

$$\text{BDI}_{\text{all},i}^{(t,r)} = \frac{\sum_{j=1}^{L} \mid C_{bias,i}^{(j;\,t,r)} \mid}{\sum_{j=1}^{L} \mid C_{bias,i}^{(j;\,t,r)} \mid + \sum_{j=1}^{L} \mid C_{feat,i}^{(j;\,t,r)} \mid + \varepsilon}.$$

This provides a compact index of the aggregate balance between feature and bias support across the network.

Competitor selection. The most generalizable form of the analysis is to estimate BDI on each input 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 the 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 the feature or by the bias component. In the special case of two classes, competitor selection is fixed by the task.

Dot-discrimination task

The convolutional neural network (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 we have previously used in humans to understand the neural correlates of decision-making (33). 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. For testing, 1,000 novel stimuli were created with equal number of red- or blue-dominant trials. All images were generated using Python.

CNN model, training and testing

We used a CNN based on the AlexNet architecture, modified for binary classification (red- versus blue-dominant images; **Fig. S1**). Convolutional-layer weights were fixed to ImageNet-1K–pretrained parameters provided by PyTorch (https://pytorch.org/), and only the fully connected layers were trained. This design keeps the feature-extraction stage stable while allowing the decision stage to adapt to the red/blue task. Using pretrained weights learned from diverse natural images also better reflects human visual development, where feature

representations are shaped by broad visual experience rather than learned from isolated red/blue dot stimuli, and it increases the likelihood that the resulting decision-composition effects generalize beyond this specific task.

The CNN was trained on randomly generated trials of the dot-discrimination 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. Training and testing were conducted through Google Colab on NVIDIA A100 GPU.

<u>Dot-discrimination task noise perturbations of input representations and readout</u>

In the dot-discrimination task, for each trial in the test set, noise was applied either to the activation pattern or to the readout weights, thereby perturbing the input 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 penultimate layer, which reflect the network's internal representation of the input. Weight noise was modeled by adding Gaussian noise, $\mathcal{N}(0, \sigma)$, to the weights mapping 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.

<u>Tiny ImageNet evaluation with ImageNet-pretrained networks</u>

To assess the generality of the analysis beyond the dot-discrimination task, we evaluated pretrained convolutional neural networks on images from the ImageNet dataset (Tiny ImageNet-200, a 200-class subset of ImageNet (35)). We analyzed all 10,000 images from the held-out validation set. 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 the 1000-way ImageNet-1K logits. When we report "label match/correctness," it denotes whether the model's top-1 predicted index equals the Tiny ImageNet class index for that image; this serves only as a post-hoc reliability annotation and is not interpreted as semantically

aligned ImageNet accuracy. Label match/mismatch refers to synset-mapped agreement between the model's top-1 ImageNet-1K prediction and the mapped ImageNet-1K label for Tiny ImageNet.

ImageNet-C evaluation under corrupted inputs

To test whether decision composition persists when feature input is degraded, we evaluated ImageNet-pretrained networks on corrupted inputs from Tiny ImageNet-C, a corruption benchmark constructed by applying ImageNet-C-style perturbations to Tiny ImageNet images (36). We focused on Gaussian noise corruptions at five severity levels (1–5). AlexNet, VGG-16, and ViT-b/16 were evaluated in inference mode using the canonical ImageNet preprocessing associated with each model's pretrained weights (resize/crop and per-channel normalization). The analysis was based on a random sample of 5,000 Tiny ImageNet images corrupted with Gaussian noise, with 1,000 images drawn at each severity level.

For each corrupted input, we defined a label-free decision margin as the difference between the highest and second-highest logits, $\Delta z = z_{(1)} - z_{(2)}$, where (1) and (2) index the model's top-1 and top-2 predicted classes for that image. Confidence was quantified as the *softmax* probability from the top-1, $confidence = softmax(z_1)$, corresponding to the probability assigned to the chosen class (neutral at 1/1000 = 0.001 for a 1000-way decision). Bias Dominance Index (BDI) at the output layer was computed in the same manner from the input-dependent feature component and an input-independent bias component.

To test whether decision composition distinguishes mechanisms beyond what is identifiable from confidence alone, we matched trials by confidence and then stratified by BDI. We stratified images into confidence bins spanning [0,1] (bin_{width} = 0.05; for a total of 20 bins) and, within each bin, split trials into low- and high-BDI subsets using a quantile criterion (bottom vs top quartile). This matched-confidence design controls for confidence by construction and isolates variation attributable to decision composition. When reporting mismatch rates, correctness (synset-mapped) was defined as whether the model's top-1 ImageNet-1K prediction matched the ImageNet-1K class index obtained by mapping the image's Tiny ImageNet WordnetID into the

ImageNet-1K synset list. This mapping is used only for reporting; BDI and confidence are computed without labels.

Confidence-BDI acceptance gating

We operationalized decision composition as a two-parameter acceptance gating policy that retains a prediction only when it is both decisive by standard uncertainty criteria and feature-grounded by the decision-composition diagnostic. For each trial i, we computed confidence, Conf_i, as the *softmax* probability assigned to the model's top-1 prediction, and computed $\mathrm{BDI}_{\mathrm{all},i}$ as the aggregate bias-dominance statistic. Given thresholds (τ, γ), we defined the accepted set as:

$$Acceptance\ (\tau, \gamma) = \{i: Conf_i \geq \tau \ \wedge \ BDI_{all,i} \leq \gamma\}$$

where τ sets a minimum confidence requirement and γ limits tolerance for bias-supported margins. Sweeping (τ, γ) over a uniform grid produced an acceptance surface whose value at each point is coverage, $\mathrm{Coverage}(\tau, \gamma) = |\ Acceptance(\tau, \gamma)\ |/N$, the fraction of predictions accepted (visualized as heatmaps with iso-coverage contours).

Data availability

Trained CNN models are available at https://osf.io/efnq6/. Tiny-ImageNet data can be downloaded at http://cs231n.stanford.edu/tiny-imagenet-200.zip.

Code availability

The analysis was based on a combination of publicly available toolboxes, analysis specific scripts. Specifically, the CNNs used in the perceptual decision-making task were based on AlexNet architecture from the PyTorch library (https://pytorch.org/). Trained AlexNet, VGG-16, and ViT-b/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/.

Acknowledgments

We would like to thank Dobromir Rahnev for his 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. K. Kassaw, F. Luzi, L. M. Collins, J. M. Malof, Are Deep Learning Models Robust to Partial Object Occlusion in Visual Recognition Tasks? *arXiv preprint arXiv:2409.10775* (2024).

2. Y. Ovadia, *et al.*, Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems* **32** (2019).

3. D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, D. Song, Natural adversarial examples in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021), pp. 15262–15271.

4. A. D'Amour, *et al.*, Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research* **23**, 1–61 (2022).

5. R. Jia, P. Liang, Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328* (2017).

6. I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

7. A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2015), pp. 427–436.

8. A. K. Menon, *et al.*, Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314* (2020).

9. B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* **30** (2017).

10. C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks in *International Conference on Machine Learning*, (2017), pp. 1321–1330.

11. R. R. Selvaraju, *et al.*, Grad-cam: Visual explanations from deep networks via gradient-based localization in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), pp. 618–626.

12. 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. (2017), pp. 3319–3328.

13. A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences in *International Conference on Machine Learning*, (2017), pp. 3145–3153.

14. S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).

15. A. Krizhevsky, One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997* (2014).

16. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

17. A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

18. S. Wang, T. Zhou, J. Bilmes, Bias also matters: Bias attribution for deep neural network explanation in *International Conference on Machine Learning*, (2019), pp. 6659–6667.

19. D. C. Knill, A. Pouget, The Bayesian brain: the role of uncertainty in neural coding and computation. *TRENDS in Neurosciences* **27**, 712–719 (2004).

20. A. Ilyas, *et al.*, Adversarial examples are not bugs, they are features. *Advances in neural information processing systems* **32** (2019).

21. R. Geirhos, *et al.*, Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**, 665–673 (2020).

22. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* **54**, 1–35 (2021).

23. S. Barocas, M. Hardt, A. Narayanan, *Fairness and machine learning: Limitations and opportunities* (MIT press, 2023).

24. M. Hort, Z. Chen, J. M. Zhang, M. Harman, F. Sarro, Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing* **1**, 1–52 (2024).

25. Y. Geifman, R. El-Yaniv, Selective classification for deep neural networks. *Advances in neural information processing systems* **30** (2017).

26. D. Kiyasseh, A. Cohen, C. Jiang, N. Altieri, A framework for evaluating clinical artificial intelligence systems without ground-truth annotations. *Nature Communications* **15**, 1808 (2024).

27. 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).

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

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

30. 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).

31. 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).

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

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

34. 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).

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

36. D. Hendrycks, T. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261* (2019).