# Using artificial neural networks to relate external sensory features to internal decisional evidence

Marshall L. Green[1], Mingjia Hu[2], Rachel N. Denison[3], Dobromir Rahnev[1]

[1]School of Psychology, Georgia Institute of Technology, Atlanta, GA

[2]Department of Psychology, Indiana University

[3]Department of Psychological and Brain Sciences, Boston University, Boston, MA

**Keywords**: perceptual decision-making, orientation discrimination, convolutional neural network

**Acknowledgement:** This work was funded by the National Institute of Health (award: R01MH119189) and the Office of Naval Research (award: N00014-20-1-2622) awarded to D.R.

**Corresponding Author:** Marshall L. Green, School of Psychology, Georgia Institute of Psychology, 654 Cherry Street Northwest, Atlanta, GA 30332. E-mail: marshall.l.green@outlook.com.

**Data availability:** Data available at https://osf.io/v6q3d/

**Competing Interest Statement:** The authors declare no competing interests.

## Abstract

All theories of perceptual decision-making postulate that external sensory information is transformed into the internal evidence that is used to guide behavior. However, the nature of this external-to-internal transformation is generally unknown. In two experiments, we examined how a particular stimulus feature – orientation – is transformed into internal evidence. Subjects judged whether Gabor patches were tilted clockwise or counterclockwise. The results of Experiment 1 demonstrated that increasing orientation offset in fine-scale increments resulted in a linear increase in sensitivity (d´), suggesting a linear external-to-internal transformation. However, the results of Experiment 2 demonstrated that increasing orientation offset in coarse-scale increments had little effect on sensitivity, suggesting a highly non-linear transformation. These behavioral results imply that a given sensory feature may not have a one-to-one mapping with the internal representation of evidence across different tasks. Further, we evaluated whether an artificial neural network (ANN) trained on orientation categorization can reproduce the observed external-to-internal transformations. The ANN mirrored the empirical results – fine-scale increments in orientation offset were linearly transformed into internal evidence, but coarse-scale increments in orientation offset had little influence on internal evidence. These results begin to reveal how external sensory information is transformed into internal decisional evidence and suggest that ANNs could serve as a hypothesis-generation platform for this critical transformation.

**Word count**: 209

## Introduction

We often use visual information to guide our behavioral decisions. The color ‘red’, for example, is used to decide whether to stop or go when driving an automobile, or similarly to select a ripe versus unripe tomato. These kinds of perceptual decisions require that we make a judgment about the identity of a stimulus based on the external sensory information that is available to us. All theories of perceptual decision-making postulate that this external sensory information is transformed into the internal evidence that is used to guide behavior. The hallmark of many theories – including signal detection theory (D. M. Green & Swets, 1966), the drift-diffusion model (Ratcliff, 1978), and ideal-observer models (e.g., Ma et al., 2006) – is the general notion that increasing the magnitude of the external feature leads to an increase in the strength of the internal evidence signal. However, the exact nature of this external-to-internal transformation remains unclear (Figure 1).

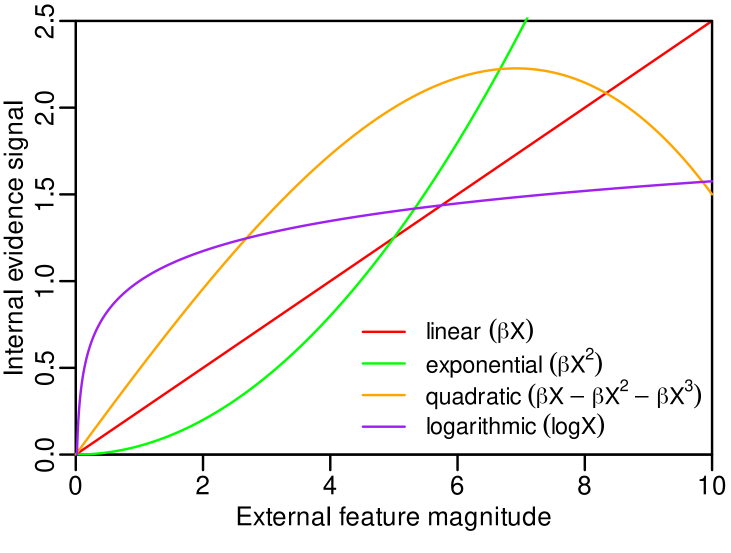

**Figure 1. Hypothetical external-to-internal transformations**. Theories of perceptual decision-making postulate that external sensory information is transformed into the internal evidence that is used to guide behavior, yet the nature of this external-to-internal transformation remains unknown. The linear model (red line) predicts a one-to-one relationship between the internal

evidence signal and the external feature. The exponential model (green line) predicts that the strength of the internal evidence signal changes at an increasing rate as the external feature is changed. The quadratic (orange line) model predicts that the internal signal strength increases before peaking and then decreasing (or increasing). The logarithmic (purple line) model predicts that the internal signal strength increases by incrementally smaller amounts (and potentially decreases) as the external signal is increased.

Most psychophysics studies do not assume a priori the function that describes how external sensory features are transformed into internal evidence. Instead, the relationship between the manipulation of any given sensory feature and performance is characterized post hoc. Although no single function is likely to universally describe the external-to-internal transformation for all perceptual domains, there have been many attempts to identify one. For example, Fechner (1860) proposed that subjective perceptual experiences were logarithmic functions of the external stimulation (Figure 1, purple line). Almost exactly one century later Stevens (1961) proposed that power functions (Figure 1, green and orange lines), rather than a logarithmic function, better describe the external-to-internal transformations of sensory systems (see also Naka & Rushton, 1966). Power functions can capture both linear relationships as well as saturation effects allowing this class of functions to flexibly fit a range of relationships between manipulations of sensory features and performance (Adler & Ma, 2018). Similarly, negative power functions do well at capturing how psychological similarity decreases exponentially as the external feature distance is increased (Shepard, 1987; Sims, 2018). Measuring the similarity function for a given feature has been useful for improving models of many cognitive processes (Nosofsky, 1992), including recent advances in visual working memory performance (Schurgin et al., 2020). However, although it seems reasonable that any given function for one feature will not describe the external-to-internal transformation for another (Augustin, 2008), it seems equally reasonable to assume that a given transformation function would hold within that feature class. Still some effects, like the oblique effect in orientation (Appelle, 1972) and motion perception (Gros et al., 1998), are examples in which the effect of some sensory features vary nonlinearly and nonmonotonically across the sensory space (Li et al., 2003). This kind of variability in how a given sensory feature affects performance leaves it unclear whether a particular external-to-internal mapping will be universal across different task contexts even for highly similar decisions within the same feature class.

Models of perceptual decision making assume different external-to-internal transformations. A straightforward model assumption is that the decision variable is a function of measurement alone, for example the strongest sensory neuron response (i.e., winner-take-all, Lee et al., 1999). This class of models assumes that sensory features have a fixed mapping with the internal decision variable, but they are agnostic to sensory uncertainty (Adler & Ma, 2018). Population coding models capture this sensory uncertainty by considering noise in sensory measurements across a population of neurons, yet the nature of the external-to-internal transformation depends on how the decision variable is computed from the sensory measurements. For example, a conventional population coding approach is to sum the logarithms of these sensory measurements to generate the log-likelihood function and then apply some algorithm to compute the decision variable. Taking the argmax of the likelihood function, a common approach which is akin to winner-take-all but in the likelihood space (Webb et al., 2007, 2010), has been demonstrated to be equivalent to a linear transformation of sensory evidence into the decision variable (Jazayeri & Movshon, 2006). Other algorithms can be applied to the likelihood function to compute the decision variable, and any linear function applied to the decision variable would simply scale or shift it in the decision space, but any nonlinear function could also be applied, for example a quadratic function. Both the linear and quadratic external-to-internal transformations can approximate a Bayesian transformation despite not being based on the Bayesian decision variable (Adler & Ma, 2018). This overall class of models has dominated recent theory development for characterizing how particular sensory features are transformed into decisions (Wohrer et al., 2013). Despite the success of many of these modeling approaches, the external-to-internal transformation is rarely established empirically beyond the scope of the model development. A general method for discovering such transformations is therefore needed.

Artificial neural networks (ANNs) are a powerful way to model human behavior (Doerig et al., 2023; Kriegeskorte, 2015; Ma & Peters, 2020). Modern deep-learning models share similarities with the connectionist networks of the 1980's and 1990's (Feldman & Ballard, 1982). However, deep-learning models differ in that they often have many layers in between the input and output layers, each of which applies a sequence of linear or nonlinear operations on the output from prior layers. An example of deep-learning ANN that is commonly used to analyze visual stimuli is convolutional neural networks which utilize convolutional layers as feature detectors to integrate information across spatially localized regions of an image (Kheradpisheh et al., 2016). Although the ANN architectures vary (e.g., VGG, ResNet, CORnet), they are the leading class of models of the mechanisms underlying primate vision (Kubilius et al., 2019). In practice, ANNs are trained to classify large sets of visual stimuli until they can accurately predict a novel visual stimulus. This learning process makes them particularly useful for discovering how sensory features map onto internal evidence because the external-to-internal mapping is emergent.

Here we empirically test how a particular stimulus feature – orientation – is transformed into internal decision evidence, and further test whether this transformation holds for similar decisions with highly different task demands. Experiment 1 demonstrated that in a high-contrast fine-discrimination task (Figure 2A), orientation is linearly transformed into internal evidence such that stimulus sensitivity is proportionate to the orientation offset of the stimuli. However, Experiment 2 showed that a low-contrast coarse-discrimination task (Figure 2B) produces a very different relationship with orientation having little influence on the internal evidence. Critically, we investigated whether an ANN could reproduce our external-to-internal mapping results. We found that an ANN trained on orientation discrimination mirrored the observed pattern of results – fine-scale increments in orientation offset were linearly transformed into internal evidence, but coarse-scale increments in orientation offset had little influence on internal evidence. These results show that a given sensory feature may not have a one-to-one mapping with the internal representation of evidence across different tasks. Critically, our findings suggest that ANNs could serve as a hypothesis-generation platform for this critical external-to-internal transformation, such that one can examine their behavior in detail, generate novel hypotheses, and then test them in human subjects.

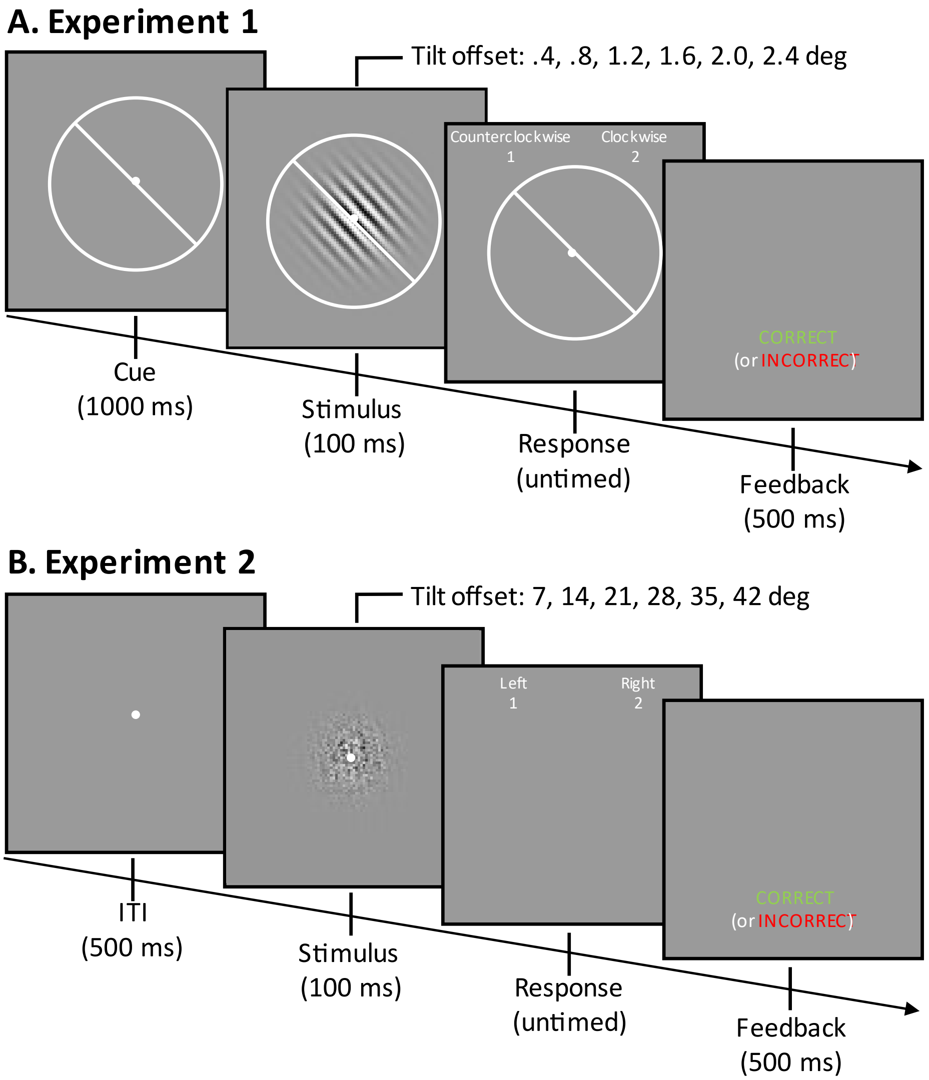

**Figure 2. Tasks for Experiments 1 and 2**. (A) Experiment 1. Subjects judged whether high-contrast Gabor patches were tilted clockwise or counterclockwise from a 45-degree cue. The Gabor patches varied in fine-scale increments from .4 to 2.4 degrees. (B) Experiment 2. Subjects judged whether noisy, low-contrast Gabor patches were tilted clockwise or counterclockwise from vertical. The Gabor patches varied in coarse-scale increments from 7 to 42 degrees.

## Methods

### *Participants*

A set of 13 subjects were recruited for Experiment 1 and a separate set of 13 subjects were recruited for Experiment 2 (26 subjects total). All subjects had normal or corrected to normal vision. One subject in Experiment 1 and one subject in Experiment 2 were excluded from analyses due to chance performance (50% accuracy across all conditions). All subjects provided informed consent and were compensated for their participation. All experimental methods and protocols were approved by the Georgia Institute of Technology Institutional Review Board.

### *Task and stimuli*

In Experiment 1, subjects performed a fine-scale orientation categorization task in which a Gabor stimulus was tilted counterclockwise or clockwise of 45° (Figure 2A). On each trial, subjects fixated on a small white dot presented at the display center. A visual cue consisting of a circle (radius = 4.5°) and a line (length = 4.5°) oriented at 45° was presented for 1000 ms. Following fixation and cueing, a Gabor patch (radius=4°) was presented for 100 ms at full contrast. Immediately after the stimulus presentation subjects provided their response by pressing “1” to respond “counterclockwise” or “2” to respond “clockwise”. After a response was made the subject was given accuracy feedback for 500 ms.

In Experiment 2, subjects performed a coarse-scale orientation categorization task in which a Gabor stimulus was tilted left (counterclockwise) or right (clockwise) of vertical (Figure 2B). On each trial, subjects fixated on a small white dot presented at the display center for 500 ms.

Following fixation, a Gabor patch (radius=4°) was presented for 100 ms at 9% contrast embedded in random pixel noise at 90% contrast. Noise was included to prevent ceiling performance and the level of noise was selected to keep performance in a range that is between chance level (50%) and perfect performance (100%). Immediately after the stimulus presentation, subjects provided their response by pressing "1" to respond "left" or "2" to respond "right". After a response was made the subject was given accuracy feedback for 500 ms.

***Procedure***

In Experiment 1, Gabor stimuli were tilted away from 45° and tilts were manipulated across five levels (.4°, .8°, 1.2°, 1.6°, 2°, and 2.4°) that were randomly chosen on each trial and counterbalanced across counterclockwise and clockwise directions. Prior to beginning the experimental runs, subjects received task instructions and then completed practice trials consisting of 20 easy trials with a 10° tilt, 40 moderately difficult trials (10 each at 5°, 3°, 2°, and .5°), and finally 20 trials with a 1.4° tilt. Subjects completed four experimental runs each consisting of six 45-trial blocks for a total of 1,080 trials.

In Experiment 2, Gabor stimuli were tilted away from vertical, and tilts were manipulated across five levels (7°, 14°, 21°, 28°, 35°, and 42°) which were randomly chosen on each trial and counterbalanced across left and right directions. Prior to beginning experimental runs, subjects received task instructions and then performed practice trials consisting of 20 easy trials with a 42° tilt at 30% contrast, 40 moderately difficult trials at decreasing tilts 35°, 28°, 14°, and 7° (10 trials each) and at 35%, 28%, 10% and .9% contrast respectively, and finally 30 trials with a 21$^{o}$ tilt at decreasing contrasts (12%, 10%, and 9%, 10 trials each). Subjects completed four experimental runs each consisting of six 45-trial blocks for a total of 1,080 trials.

### *Apparatus*

Both Experiments 1 and 2 were designed in the MATLAB environment using Psychtoolbox 3 (Brainard, 1997). Stimuli were presented on a 21.5-inch iMac monitor (1920 × 1080 pixel resolution, 60 Hz refresh rate) in a dark room. Subjects were seated 60 cm away from the display and provided their responses using a standard computer keyboard.

### *Analysis*

To compare subjects' sensitivity to orientation across conditions, we computed d' using the standard formula (D. M. Green & Swets, 1966) by treating clockwise tilt trials as the target and calculating the hit rate (HR) and false alarm rate (FAR):

$$d' = \Phi^{-1}(HR) - \Phi^{-1}(FAR)$$

where $\Phi^{-1}$ is the inverse of the cumulative standard normal distribution that transforms the HR and FAR into z-scores. Sensitivity (d') was computed separately for each individual and each condition. This same approach was used to compute sensitivity (d´) from the HR and FAR produced by the ANN model.

We used mixed effects regression models to estimate how external sensory signal strength – the degree of orientation offset – is transformed into internal evidence as measured with sensitivity (d’) while accounting for random effects arising from individual subjects. A model which describes a linear relationship between orientation offset and sensitivity is given as:

$$d' = \alpha + \beta X + \varepsilon$$

where $\alpha$ is the intercept, $\beta$ is the slope, and $\varepsilon$ is random error.

Neither Experiment 1 nor 2 included a condition in which orientation offset was zero since such a condition would not have a correct decision. However, we assumed a hypothetical sensitivity (d’) of zero by constraining the intercept to the origin. By fixing the intercept to be zero, the intercept term can be dropped, and the linear model can be reduced to:

$$d' = \beta X + \varepsilon$$

To account for potential nonlinear external-to-internal mappings, we fit several additional models. First, we fit a quadratic regression model whereby the effect of manipulating orientation offset on sensitivity (d’) is given as:

$$d' = \beta_1 X + \ \beta_2 X^2 + \varepsilon$$

We then fit a polynomial regression model by including an additional exponent term of the third degree whereby the effect of manipulating orientation offset on sensitivity is given as:

$$d' = \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

Finally, we fit a logarithmic model whereby the effect of manipulating orientation offset on sensitivity is given as:

$$d' = \beta_1 X + \beta_2 logX + \varepsilon$$

To assess which model best describes the transformation of external sensory signals into internal evidence, we used Bayesian Information Criterion (BIC; Schwarz, 1978) to compare the relative fits of these regression models to each other. BIC penalizes model complexity by taking the product of the number of parameters and the natural log of the number of data points. Similar results are found if models are compared using Akaike’s information criterion (AIC; Akaike, 1973) which is more lenient on model complexity than BIC because the penalty term is a constant factor of two.

### *Artificial neural network (ANN) model*

We built a simple feedforward ANN model (Figure 3) using the TensorFlow toolbox (Abadi et al., 2016). The model takes as input a 100×100 image and outputs one of two category labels corresponding to clockwise or counterclockwise orientation offset. The model consists of one convolutional layer, one pooling layer, and a fully connected layer. The convolutional layer consists of a set of four linear filters with equally sized receptive fields (3×3 pixels) with equally spaced intervals followed by sigmoid activation. This processing operation results in a 98×98 activation map encoding the response of a given filter at each spatially localized regions of an image. The activation maps of all four filters within the convolutional layer are stacked to produce a 98×98×4 output volume. The output of the convolutional layer is pooled to reduce the size of the input volume by taking the maximum activation value of 2×2 spatially localized units resulting in a 49×49×4 output volume. This pooled result was then flattened and fully connected to two decision units. This choice of hyperparameters and stimulus size resulted in 19,250 trainable weights.

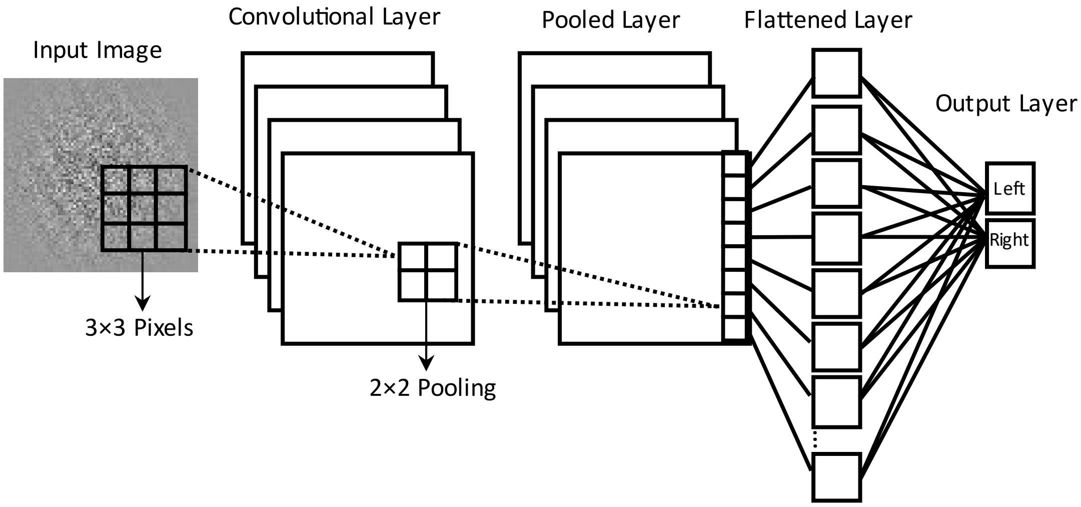

**Figure 3. Schematic of the ANN model**. The model takes as input a 100×100 image and outputs one of two category labels corresponding to counterclockwise (left) or clockwise (right) orientation. The model consists of one convolutional layer, one pooling layer, and a fully connected layer. The convolutional layer consists of a set of four linear filters with equally sized receptive fields (e.g., 3×3 pixels) with equally spaced intervals followed by sigmoid activation. The output of the convolutional layer is pooled to reduce the size of the input volume by taking the maximum activation value of 2×2 spatially localized units. This pooled result was then flattened and fully connected to two decision units.

The model was trained to categorize Gabor stimuli as tilted leftward or rightward of vertical. The total training set consisted of 10,000 stimuli. During training, the orientation offset, contrast, noise, and phase of each Gabor stimulus was randomly drawn from uniform distributions. The sampling distribution for orientation offset ranged from 0 to 45 degrees, for contrast from 1% to 90% (i.e., amplitude from .01 to .9) and for noise from 1% to 100%. Visual noise was created by embedding random pixel noise into the Gabor stimulus. The midpoint gray of the stimulus was defined as 0 with black being -1 and white being 1. Pixel values less than -1 were set equal to -1, and pixels greater than 1 were set equal to 1, ensuring that the total stimulus contrast was within the range of 0 to 100%. Learning was evaluated by testing the model on 1,000 novel Gabor stimuli generated from the same sampling distributions for orientation offset, contrast, noise, and phase. The model accurately categorized the orientation of 99 percent of the validation stimulus set following 10 training epochs, showing that the model successfully learned the orientation categorization task.

The trained ANN was then tested on fine and coarse-scale orientation categorization tasks. For the fine-scale orientation categorization task, orientation offset was varied across 20 equally spaced levels ranging from 0 to 2.8 degrees. For the coarse-scale orientation categorization task, orientation offset was varied across 20 equally spaced levels ranging from 0 to 42 degrees. For both the fine- and coarse-scale tests of the ANN, the Gabor stimuli varied in contrast across five levels ranging from 3% to 9%. Each test stimulus was presented at 100% visual noise, and the phase of each Gabor wavelet was randomly sampled uniformly from 0 and 1. The model was tested on 1,000 stimuli at each combination of orientation offset and contrast, resulting in a total of 100,000 simulated trials. To reduce the chance of idiosyncratic model behavior due to the random starting weights, we trained 30 model initializations of the model and averaged their results.

### *Data and code*

The data and analyses code are available at https://osf.io/v6q3d/.

## Results

Our goal was to examine how external sensory information is transformed into internal decisional evidence and examine whether ANNs can serve as a hypothesis-generation platform for this critical transformation. In two behavioral experiments, we empirically tested how a particular stimulus feature – orientation – is transformed into internal evidence. Subjects judged whether a Gabor stimulus was tilted clockwise or counterclockwise from a criterion. Across the two experiments we tested how orientation is transformed into internal evidence for similar decisions with highly different task demands.

### *Behavioral results*

In Experiment 1, subjects judged the orientation offset of a high-contrast Gabor that was tilted in fine-scale increments from .4 to 2.4 degrees away from a 45-degree cue (Figure 2A). We found that sensitivity (d') appears to increase linearly as tilt was increased (Figure 4A). To examine the linearity of the function, we fit hierarchical regression models to characterize how sensitivity (d') varied as a function of orientation offset with the subject factor as the random effect variable. We found that a slope-only linear model (where the intercept was constrained to zero) provided the best fit to the data (best fitting model: $d' = .72x$; Figure 4B). Indeed, the linear model outperformed the quadratic ($\Delta BIC = 16.31, \Delta AIC = 9.25$), third-degree polynomial ($\Delta BIC = 25.07, \Delta AIC = 8.57$), and logarithmic models ($\Delta BIC = 15.98, \Delta AIC = 8.91$). In addition, we fit a full linear model ($\Delta BIC = 16.35, \Delta AIC = 9.28$; model: $d' = -.03 + .74x$) that includes both a slope and an intercept and found that the intercept was not significantly different from zero ($\alpha = -0.03, SE = .07, t(64) = -0.47, p = .64$). Although each of the models can reasonably capture the data, suggesting that these models are not well differentiated by the fine-scale task, the linear model with the intercept constrained to zero provides the most parsimonious fit to the data. Altogether, this pattern of results suggests that orientation is linearly transformed into internal decision evidence.

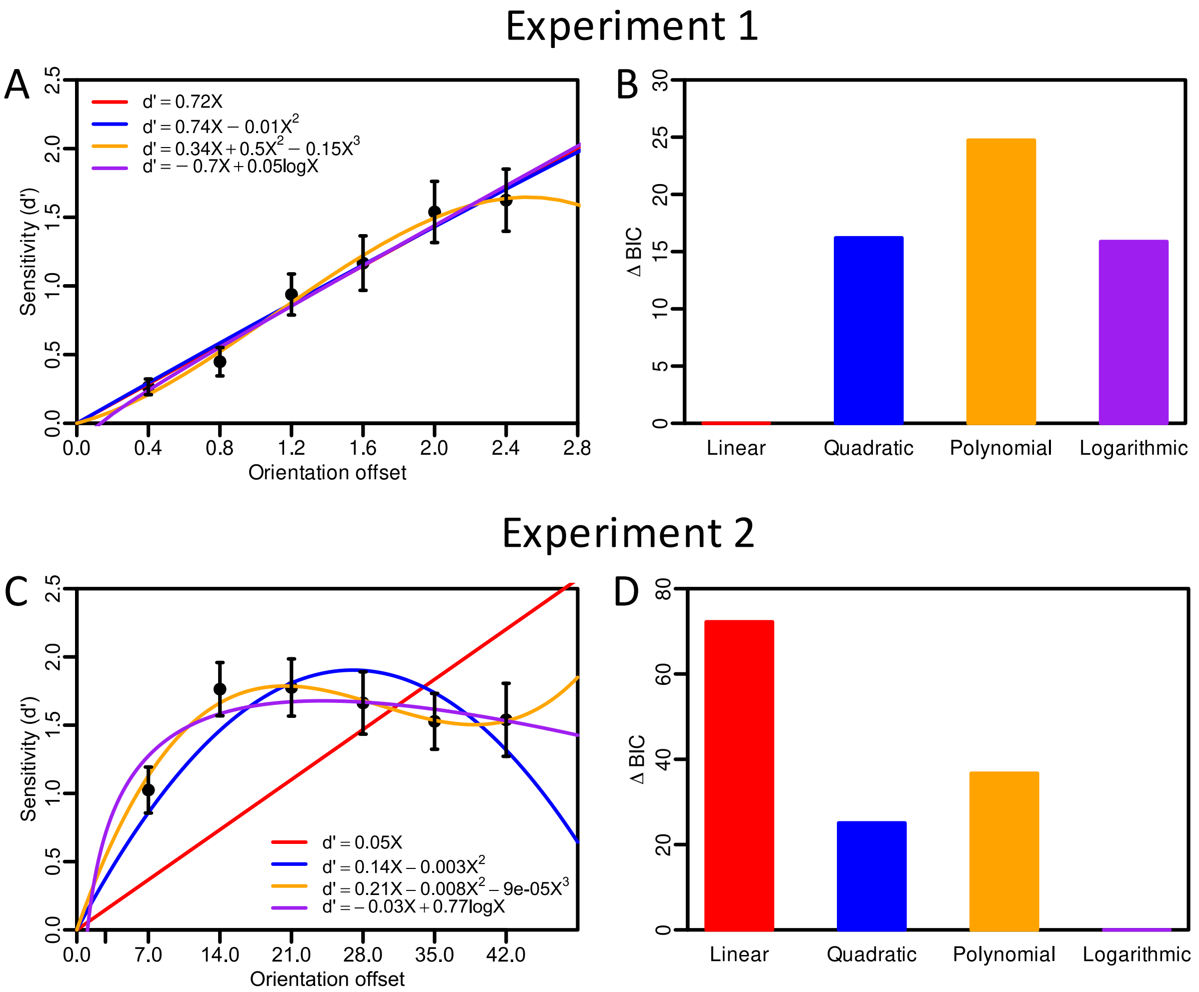

**Figure 4. Experiment 1 and 2 behavioral results**. (A) The results of Experiment 1 show that increasing orientation offset in fine-scale increments results in a linear increase in sensitivity (d'). (B) A model comparison using BIC demonstrates that the linear model with the intercept constrained to the origin (red line) was the best performing model, suggesting a linear external-to-internal transformation. (C) The results of Experiment 2 show that sufficiently large increases in orientation offset had little effect on sensitivity. (D) BIC suggested that the logarithmic model

(purple line) was the best performing model, suggesting a highly non-linear transformation. Error bars in panels A and C show SEM.

Critically, we found a strikingly different pattern of results for Experiment 2 where subjects judged the orientation offset of a noisy, low-contrast Gabor in coarse-scale increments from 7 to 42 degrees away from vertical (Figure 2B). Instead of sensitivity (d') linearly increasing with the magnitude of orientation offset, we found that above 14 degrees tilt sensitivity no longer increased with orientation offset (Figure 4C). This effect was not due to either floor or ceiling effects as the accuracy in the different conditions was in a range ($69 - 79\%$) that is far from both chance level (50%) or perfect performance (100%). After 14 degrees there was a numerical decrease in orientation sensitivity that is, however, not significant (*p*s>=.41).

This pattern of results was best fit by a logarithmic model that features a steep initial increase followed by a mostly flat (and, in fact, slightly decreasing) portion. The logarithmic model outperformed a slope-only linear model ($\Delta BIC = 71.60, \Delta AIC = 78.17$), a full linear model ($\Delta BIC = 15.83, \Delta AIC = 15.84$), a quadratic model ($\Delta BIC = 25.54, \Delta AIC = 25.54$), and a polynomial model ($\Delta BIC = 124.46, \Delta AIC = 115.70$). Taken together, these results reveal that, under the task demands of Experiment 2, orientation is strongly nonlinearly transformed into internal decision evidence.

***ANN results***

The results from Experiments 1 and 2 show that intuitively similar external stimulus features can have extremely different mappings to internal evidence. While these results are likely to be explainable within different modeling frameworks, it is not clear whether any model frameworks would have predicted them a priori without additional assumptions about perceptual similarity functions. For example, it appears that an off-the-shelf probabilistic population coding model does not predict these results (see Supplementary), though it is of course likely that a better fit can be obtained if additional assumptions are included in the model.

Here we evaluated whether an ANN model trained on orientation categorization would naturally reproduce the observed external-to-internal transformations found in Experiments 1 and 2 without any additional assumptions or training. We trained a 2-layer ANN model on discriminating between Gabor patches tilted clockwise or counterclockwise from vertical. To provide an unbiased training set, we trained the ANN on a wide range of tilts (0 to 45 degrees) and contrasts (1 to 90%). We then tested the trained ANN on stimuli that mimic Experiment 1 (fine-grained tilts up to 2.8 degrees) and Experiment 2 (coarse-grained tilts up to 42 degrees). We found that the ANN model reproduced both the linear relationship between sensitivity and orientation for fine-grained tilts that we observed in Experiment 1 (Figure 5A), and an increase-then-plateau relationship between sensitivity and orientation for coarse-grained tilts that we observed in Experiment 2 (Figure 5B). To check for the robustness of these effects, we examined them for different contrast levels from 3 to 9% contrast and found that both patterns remained the same regardless of contrast (Figure 5A,B). Nevertheless, unlike the human data in Experiment 2 where d’ peaks around 14 degrees, the ANN model has maximum d’ at around 8 degrees (see Discussion). These results suggest that the ANN model shows very similar emergent behavior to what we see in humans for the transformation from external sensory features to internal decision evidence.

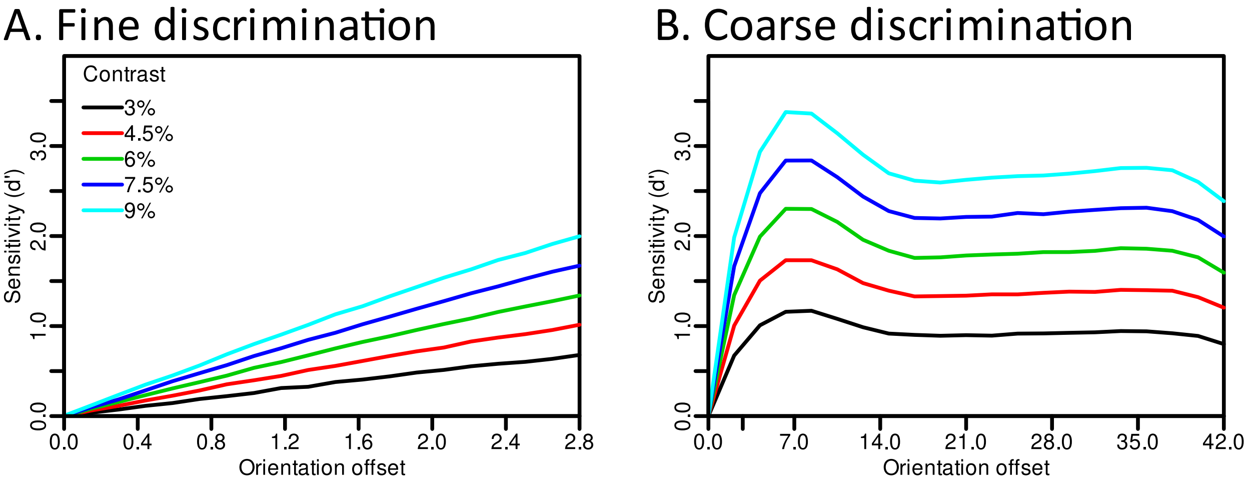

**Figure 5. Artificial Neural Network (ANN) results**. The ANN reproduced the empirical results – fine-scale increments in orientation offset were linearly transformed into internal evidence (A), but coarse-scale increments in orientation offset were nonlinearly transformed into internal evidence (B).

## Discussion

We examined how external sensory information is transformed into internal decisional evidence across different contexts. In Experiment 1, increasing the orientation offset of a high-contrast Gabor in fine-scale increments away from 45 degrees resulted in a linear increase in sensitivity (d'), suggesting a linear transformation from orientation to internal evidence strength. In contrast, in Experiment 2, increasing the orientation offset of a noisy, low-contrast Gabor in coarse-scale increments away from vertical resulted in a fast initial gain in sensitivity followed by a plateau, suggesting a highly non-linear relationship between orientation and internal evidence strength. Critically, an artificial neural network (ANN) model trained on the orientation discrimination task reproduced the observed pattern of results. These results demonstrate that the task context can dramatically change how sensory information is transformed into internal decisional evidence within the same visual domain and suggest that ANNs can serve as a hypothesis-generation platform for this critical transformation.

### *Dissociation between tasks using fine- and coarse-scale stimuli*

The general intuition of most theories of perceptual decision-making is that increasing the strength of a given external signal leads to a graded increase in the strength of the internal signal (e.g., Green & Swets, 1966; Ma et al., 2006; Ratcliff, 1978). In contrast with this intuition, we show that in the task with coarse-scale stimuli (Experiment 2) a large increase of the external signal does not translate into an increase in the internal signal. Crucially, this pattern does not appear to reflect a ceiling or a floor effect because performance saturates in a range between 69 and 79% accuracy. What explains this pattern of results? There appear to be at least three different explanations.

First, it could be that the results are explained by the fact that the optimal readout mechanisms differ between the fine- and coarse-scale discrimination tasks. Fine-scale tasks require a subject to discriminate between signals that are nearby in the stimulus space, whereas coarse-scale tasks require a subject to discriminate between far apart signals. Performance differences across fine- and coarse-scale tasks have been found to be a function of the activity of feature-selective neurons (Britten et al., 1992; Celebrini & Newsome, 1995; Salzman et al., 1990, 1992). Optimal performance depends on the readout mechanisms from these neurons – for fine-scale tasks, similar stimulus features activate roughly the same population of feature selective neurons and so the most informative neurons are those tuned slightly away from the feature to be discriminated (Jazayeri & Movshon, 2007; Scolari & Serences, 2010; Verghese et al., 2012). For coarse-scale tasks, the most informative neurons are those tuned to the to-be-discriminated feature. It could be that similar mechanisms are involved in fine and coarse orientation categorization tasks whereby sufficient increases in orientation offset changes the informative value of responding neurons and optimal performance requires transitioning from a fine-scale scheme to a coarse-scale scheme. However, whereas these different decision mechanisms are likely responsible for some aspects of the differing patterns of performance across the fine- and coarse-scale tasks, they do not directly explain why performance plateaus in the coarse-scale task.

Second, it could be that the coarse-scale task involves threshold mechanisms not present in the fine-scale tasks. Specifically, it may be that in the coarse-scale task, which involves noisy Gabor patches, there is some intensity of the external feature needed to perceive the stimulus (Rouder & Morey, 2009). Crucially, in the coarse-scale task, signal-to-noise ratio (SNR) is still increased across conditions because the Gabor stimuli in each orientation condition contains the same amount of visual noise, and increasing the SNR is expected to result in some amount of increase in orientation sensitivity. Instead, we found that a sufficiently large increase in SNR had no additional effect on sensitivity. In other words, this flat performance suggests that the high amount of visual noise (90%) used in the coarse discrimination task may have functioned as a threshold on identifying the orientation of the Gabor whereby the observer either perceives the orientation signal and can easily categorize whether it is tilted left or right, or the observer fails to perceive the orientation signal and cannot identify the orientation at all regardless of the magnitude of the orientation offset. This explanation is in line with recent work demonstrating that different mechanisms can result in either graded or all-or-none perception even for highly similar visual stimuli (M. L. Green & Pratte, 2022).

Third, rather than stemming from different decisional or threshold mechanisms, it may simply be that any external feature is transformed into internal evidence in complex ways that depend on a host of factors and are difficult to intuit. For example, although we see very different patterns of results across Experiments 1 and 2, it could be that the pattern across these two experiments is explained by an underlying relationship between perceptual similarity and feature distance which follows the Weber-Fechner law. A small increase in the difference between the feature and the decision boundary (e.g., vertical) may initially result in an increase in perceptual sensitivity for detecting that difference, but with little-to-no increase in sensitivity for detecting sufficiently large feature distances. However, given the other differences between the tasks in Experiment 1 and 2, including visual noise, contrast, and the decision boundary, it is not clear whether larger increases of orientation offset in Experiment 1, or conversely smaller orientation offsets in Experiment 2, would follow the Weber-Fechner law. Any of these factors could increase the complexity with which the external feature is transformed into internal evidence. According to this possibility, having a complete model of our visual system would allow us to discover many such complex relationships, but at least some of them likely will not have a succinct and intuitive explanation. The two conditions examined here differed in the task being coarse- vs. fine-scale, but also in the stimulus contrast and stimulus noise, with each of these factors possibly having a difficult-to-predict influence. This possibility does not allow us to predict behavioral patterns a priori without a model of the visual system, but it may allow this in the presence of such a model.

Although the ANN reproduced the overall pattern of results across the fine- and coarse-scale orientation categorization tasks, there was a notable difference between the ANN and human data. Unlike the human data in Experiment 2 where d’ peaks around 14 degrees, the ANN model has maximum d’ at around 8 degrees which then decreases before saturating. Overall, manipulating stimulus contrast had the effect of scaling model performance, but whereas the peak and decrease is less pronounced at lower contrast and more exaggerated at higher contrast, the location of the peak did not change. We attempted to identify what might be causing this pattern of results in the ANN's task performance. Our first intuition is that the pattern was driven by how much noise there is in the stimulus. However, like the effect of manipulating contrast, manipulating the amount of noise only scaled performance without affecting the peak. Another possibility was that this pattern reflected idiosyncratic behavior arising from random initialization of model weights, but we reject this explanation because each of the thirty-model initialization reflected a highly similar pattern. A third possibility that we did not explore is whether the pattern is caused by our choice of hyperparameters, such as the choice of using a 3 × 3-pixel receptive field size. Although it is unclear at this point what drives this slight difference in performance between humans and the ANN model, we posit that identifying what causes the model to exhibit this pattern of results will generate a testable hypothesis for future experiments with human subjects.

### ***Using ANNs as hypothesis-generation platforms***

One of the big promises of ANN models is that they can function as increasingly more appropriate models of the human visual system (Doerig et al., 2023; Kriegeskorte, 2015). It is clear that current versions of these models differ from human visual perception in many ways (Bowers et al., 2022), which is not surprising given the vast differences between brains and ANNs in both architecture and training. Nevertheless, despite these vast differences, many similarities between ANNs and brains have also been reported (Kheradpisheh et al., 2016; Kubilius et al., 2019). The existence of these similarities suggests that ANNs may sometimes be useful as hypothesis-generation platforms even without aligning their architecture or training with that of human brains. The reasoning here is that some tasks may involve built-in constraints, such that most systems that learn to complete the task, regardless of their details, will also exhibit the same dependencies. We believe that this may be why the simple ANN used here was able to reproduce, out of the box, the complex qualitative pattern in human data despite this ANN being so different from human brains. If so, many ANNs may already be useful as hypothesis-generation platforms, at least in the specific cases where they are trained and tested on the same specific task performed by human subjects. On the other hand, when ANNs are trained on one task/dimension and tested on a different task/dimension, there is little reason to believe that they will behave similarly to humans.

Here we tested how a very simple visual feature, orientation, maps onto internal evidence in the context of a single class of stimuli (Gabor patches). The simplicity of this task allowed us to build a relatively small and shallow ANN (with just two layers). In fact, the simplicity of the task makes it superfluous to employ a deep network, such as the ones in most contemporary deep-learning models (e.g., VGG or ResNet). Indeed, we found that even this simple ANN model mirrored human performance without additional assumptions or special training. However, more complex features, such as ones that allow for image classification or person recognition, will certainly require deeper and more complex networks. Thus, the network chosen as a hypothesis-generation platform should have complexity commensurate to the task at hand.

### ***Conclusion***

Whereas previous work has shown that the external-to-internal mapping often varies from one visual domain to another, here we show that the mapping varies drastically across tasks within a visual domain. We further demonstrated that a shallow ANN, trained on the orientation discrimination task, mirrored the pattern of results observed in human subjects without any additional assumptions or training. Taken together, these results begin to reveal how external sensory information is mapped onto internal decisional evidence. Critically, our findings suggest that artificial neural networks could serve as a powerful hypothesis-generation platform for building a theory of this critical external-to-internal transformation, such that one can examine their behavior in detail, generate novel hypotheses, and then test them in human subjects.

## References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv Preprint arXiv:1603.04467*.

Adler, W. T., & Ma, W. J. (2018). Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLOS Computational Biology*, *14*(11), e1006572. https://doi.org/10.1371/journal.pcbi.1006572

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Technology*, 199–213.

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: The "oblique effect" in man and animals. *Psychological Bulletin*, *78*(4), 266–278. https://doi.org/10.1037/h0033117

Augustin, T. (2008). Stevens' power law and the problem of meaningfulness. *Acta Psychologica*, *128*(1), 176–185. https://doi.org/10.1016/j.actpsy.2007.12.005

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2022). Deep Problems with Neural Network Models of Human Vision. *Behavioral and Brain Sciences*, 1–74. https://doi.org/10.1017/S0140525X22002813

Brainard, D. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Britten, K., Shadlen, M., Newsome, W., & Movshon, J. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. *The Journal of Neuroscience*, *12*(12), 4745–4765. https://doi.org/10.1523/JNEUROSCI.12-12-04745.1992

Celebrini, S., & Newsome, W. T. (1995). Microstimulation of extrastriate area MST influences performance on a direction discrimination task. *Journal of Neurophysiology*, *73*(2), 437–448. https://doi.org/10.1152/jn.1995.73.2.437

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., Van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*(7), 431–450. https://doi.org/10.1038/s41583-023-00705-w

Fechner, G. T. (1860). Elemente der Psychophysik (Reissued 1964 by Bonset, Amsterdam ed.). *Leipsiz: Breitkopf & Hartel*.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist Models and Their Properties. *Cognitive Science*, *6*(3), 205–254. https://doi.org/10.1207/s15516709cog0603_1

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). Wiley.

Green, M. L., & Pratte, M. S. (2022). Local motion pooling is continuous, global motion perception is discrete. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(1), 52–63. https://doi.org/10.1037/xhp0000971

Gros, B. L., Blake, R., & Hiris, E. (1998). Anisotropies in visual motion perception: A fresh look. *Journal of the Optical Society of America A*, *15*(8), 2003. https://doi.org/10.1364/JOSAA.15.002003

Jazayeri, M., & Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience*, *9*(5), 690–696. https://doi.org/10.1038/nn1691

Jazayeri, M., & Movshon, J. A. (2007). A new perceptual illusion reveals mechanisms of sensory decoding. *Nature*, *446*(7138), 912–915. https://doi.org/10.1038/nature05739

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. *Scientific Reports*, *6*(1), 32672. https://doi.org/10.1038/srep32672

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. *Advances in Neural Information Processing Systems*, *32*.

Lee, D. K., Itti, L., Koch, C., & Braun, J. (1999). Attention activates winner-take-all competition among visual filters. *Nature Neuroscience*, *2*(4), 375–381. https://doi.org/10.1038/7286

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique Effect: A Neural Basis in the Visual Cortex. *Journal of Neurophysiology*, *90*(1), 204–217. https://doi.org/10.1152/jn.00954.2002

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438. https://doi.org/10.1038/nn1790

Ma, W. J., & Peters, B. (2020). *A neural network walks into a lab: Towards using deep nets as models for human behavior* (arXiv:2005.02181). arXiv. http://arxiv.org/abs/2005.02181

Naka, K. I., & Rushton, W. A. H. (1966). S-potentials from luminosity units in the retina of fish (Cyprinidae). *The Journal of Physiology*, *185*(3), 587–599. https://doi.org/10.1113/jphysiol.1966.sp008003

Nosofsky, R. M. (1992). Similarity Scaling and Cognitive Process Models. *Annual Review of Psychology*, *43*, 25–53.

Ratcliff, R. (1978). A Theory of Memory Retrieval. *Psychological Review*, *85*(2), 59–108. https://psycnet.apa.org/doi/10.1037/0033-295X.85.2.59

Rouder, J. N., & Morey, R. D. (2009). The nature of psychological thresholds. *Psychological Review*, *116*(3), 655–660. https://doi.org/10.1037/a0016413

Salzman, C., Britten, K. H., & Newsome, W. T. (1990). Cortical microstimulation influences perceptual judgements of motion direction. *Nature*, *346*(6280), 174–177. https://doi.org/10.1038/346174a0

Salzman, C., Murasugi, C., Britten, K., & Newsome, W. (1992). Microstimulation in visual area MT: Effects on direction discrimination performance. *The Journal of Neuroscience*, *12*(6), 2331–2355. https://doi.org/10.1523/JNEUROSCI.12-06-02331.1992

Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*(11), 1156–1172. https://doi.org/10.1038/s41562-020-00938-0

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Scolari, M., & Serences, J. T. (2010). Basing Perceptual Decisions on the Most Informative Sensory Neurons. *Journal of Neurophysiology*, *104*(4), 2266–2273. https://doi.org/10.1152/jn.00273.2010

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656. https://doi.org/10.1126/science.aaq1118

Stevens, S. S. (1961). To Honor Fechner and Repeal His Law: A power function, not a log function, describes the operating characteristic of a sensory system. *Science*, *133*(3446), 80–86. https://doi.org/10.1126/science.133.3446.80

Verghese, P., Kim, Y.-J., & Wade, A. R. (2012). Attention Selects Informative Neural Populations in Human V1. *The Journal of Neuroscience*, *32*(46), 16379–16390. https://doi.org/10.1523/JNEUROSCI.1174-12.2012

Webb, B. S., Ledgeway, T., & McGraw, P. V. (2007). Cortical pooling algorithms for judging global motion direction. *Proceedings of the National Academy of Sciences*, *104*(9), 3532–3537. https://doi.org/10.1073/pnas.0611288104

Webb, B. S., Ledgeway, T., & McGraw, P. V. (2010). Relating spatial and temporal orientation pooling to population decoding solutions in human vision. *Vision Research*, *50*(22), 2274–2283. https://doi.org/10.1016/j.visres.2010.04.019

Wohrer, A., Humphries, M. D., & Machens, C. K. (2013). Population-wide distributions of neural activity during perceptual decision-making. *Progress in Neurobiology*, *103*, 156–193. https://doi.org/10.1016/j.pneurobio.2012.09.004

# Supplementary Materials

Our goal was to examine the predictions of an off-the-shelf probabilistic population coding model for the fine- and coarse-scale orientation discrimination tasks. Data were simulated from a model based on an encoder-decoder framework (M. L. Green & Pratte, 2022; Jazayeri & Movshon, 2006; Webb et al., 2007) in which evidence for the orientation category of a given stimulus is represented across a bank of orientation selective channels. The orientation sensitivity function of each detector followed a von Mises distribution to ensure that response profiles respected the circular nature of orientation space. The precision of each orientation detector ($\kappa$≈2.95) was chosen to approximate the Gaussian half-width half-max of $40^{0}$. Each model included 180 motion detectors centered one degree apart at $\vartheta_{\mathrm{i}}$. The sensitivity function of the *i*th orientation detector ($S_i$) to orientation $\vartheta_j$ follows:

$$S_i(\theta_j) = \frac{e^{\kappa \cos{(\theta_j - \theta_i)}}}{2\pi I_0(\kappa)}$$

The response profile of the *i*th orientation detector to a particular stimulus ($R(D)$) is given by:

$$R(D) = S_i(\theta)bg$$

where *b* is the baseline firing rate in spikes per second (10 spike/s) and *g* is response gain representing the contrast of the Gabor wavelet ($g_s$) and visual noise ($g_n$). The number of spikes from an orientation detector ($n_i$) follows a Poisson distribution with mean determined by that detector's response profile ($R_i(D)$),

$$Poisson(n_i|D) = e^{-R(D)}\frac{R(D)^{n_i}}{n_i!}$$

The distribution of orientation channel spikes is then multiplied by the log of the channel sensitivity functions to read out the evidence for a given stimulus.

$$logL\left(\theta_j\right) = \sum_{j=-90}^{1801} Poisson(n_i|D)\log\left(S_i(\theta_j)\right)$$

Evidence for whether a given stimulus was left or right of vertical (zero degrees) was computed by taking the argmax of the log Likelihood for orientations greater than zero (rightwards tilts) and less than zero (leftwards tilts). The ratio of evidence for one choice and the other is compared to the criterion of zero.

We examined how the model's orientation sensitivity changes with manipulations of orientation offset. Responses to simulated orientation stimuli were generated from each model 1000 times for each stimulus tilt and contrast. We computed d' using the standard formula (D. M. Green & Swets, 1966) by treating clockwise tilt stimuli as the target and calculating the hit rate (HR) and false alarm rate (FAR). The PPC model predicts that increasing orientation offset in fine-scale increments results in a linear increase in sensitivity (d'), a pattern that is identical to that of human subjects. Critically, the PPC model also predicts that increasing orientation offset in coarse-scale increments similarly results in a linear increase in sensitivity (d'), a pattern which is starkly different from that of human subjects. Although this off-the-shelf PPC model does not predict the same pattern of results obtained from human subjects, it is of course likely that a better fit can be obtained if additional assumptions are included in the model.

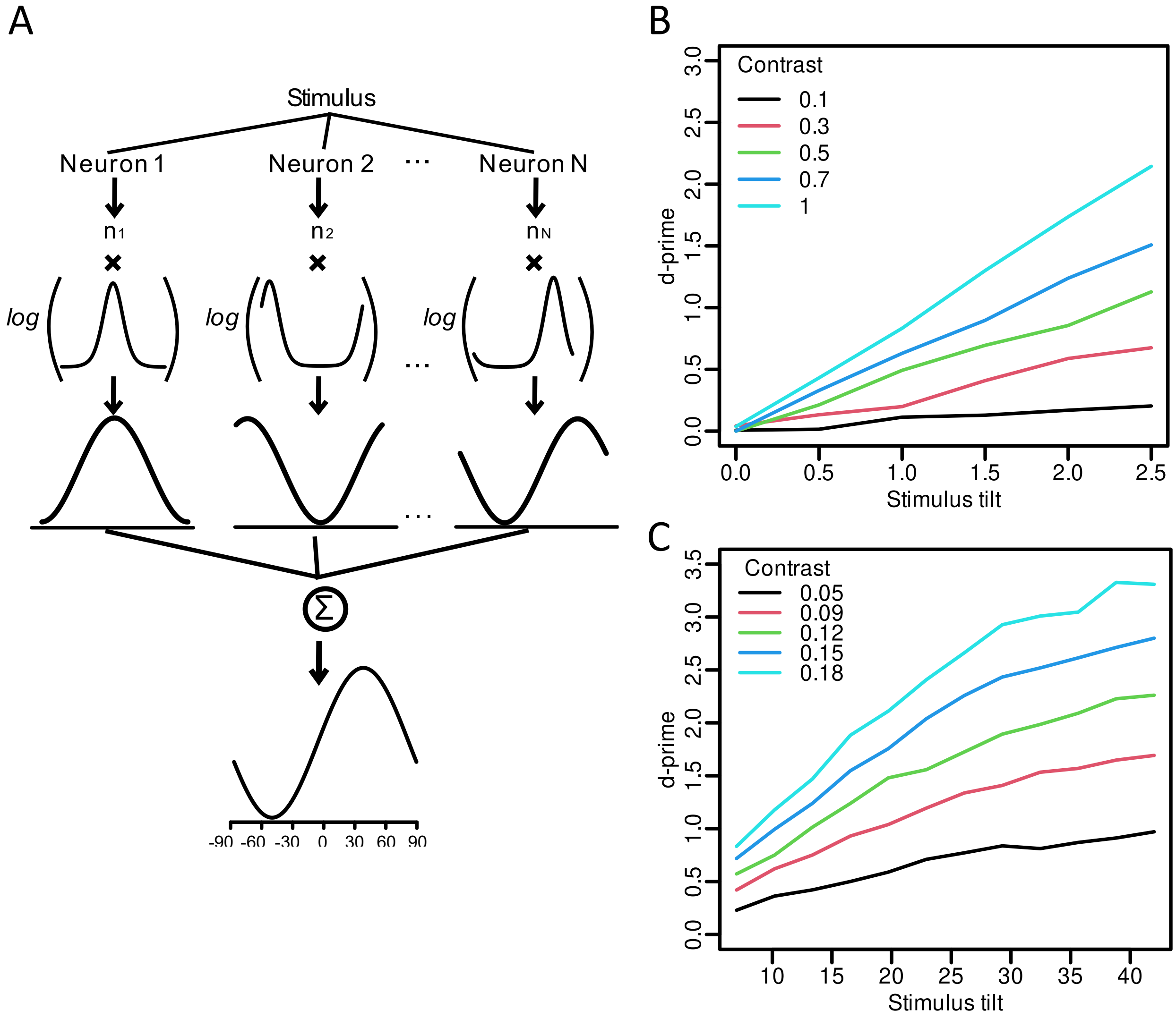

**Figure S1.** *Probabilistic Population Coding (PPC) model and synthetic data.* (A) An off-the-shelf probabilistic population coding model represents orientation evidence as a log likelihood function. (B) The PPC model predicts that increasing orientation offset in fine-scale increments results in a linear increase in sensitivity (d’). (C) The PPC model similarly predicts that increasing orientation offset in coarse-scale increments results in a linear increase in sensitivity.