Neuroscience

An image-computable model of speeded decision-making

Paul I Jaffe ✉, Gustavo X Santiago-Reyes, Robert J Schafer, Patrick G Bissett, Russell A Poldrack

Department of Psychology, Stanford University, Stanford, United States • Department of Bioengineering, Stanford University, Stanford, United States • Lumos Labs, San Francisco, United States

Reviewed Preprint
v2 • January 29, 2025
Revised by authors

Reviewed Preprint
v1 • June 13, 2024

eLife Assessment

This **important** study presents an original and promising approach to combine convolutional neural networks of visual processing with evidence accumulation models of decision-making. While the methodological approach is technically sophisticated and the evidence is **solid**, there is still a gap between the model and the behavioral data. The study will be of interest to researchers working in the fields of machine learning and cognitive modeling.

https://doi.org/10.7554/eLife.98351.2.sa3

Abstract

Evidence accumulation models (EAMs) are the dominant framework for modeling response time (RT) data from speeded decision-making tasks. While providing a good quantitative description of RT data in terms of abstract perceptual representations, EAMs do not explain how the visual system extracts these representations in the first place. To address this limitation, we introduce the visual accumulator model (VAM), in which convolutional neural network models of visual processing and traditional EAMs are jointly fitted to trial-level RTs and raw (pixel-space) visual stimuli from individual subjects in a unified Bayesian framework. Models fitted to large-scale cognitive training data from a stylized flanker task captured individual differences in congruency effects, RTs, and accuracy. We find evidence that the selection of task-relevant information occurs through the orthogonalization of relevant and irrelevant representations, demonstrating how our framework can be used to relate visual representations to behavioral outputs. Together, our work provides a probabilistic framework for both constraining neural network models of vision with behavioral data and studying how the visual system extracts representations that guide decisions.

Introduction

Decision-making under time pressure is deeply embedded within the activities of daily life. To study the cognitive and neural processes underlying decision-making, psychologists fit computational models to response time (RT) data gathered from relatively simple cognitive tasks.

Evidence accumulation models (EAMs), such as the diffusion-decision model [66], [67] and linear ballistic accumulator model (LBA) [9], are the most successful and widely-used computational models of joint RT and choice data from decision-making tasks. In the EAM framework, sensory evidence is accumulated (possibly with noise) until reaching a threshold, at which point an overt response is generated. As such, decision-making is described in terms of a small number of interpretable parameters that capture basic cognitive processes. Empirically, EAMs have been shown to capture RT distributions from a variety of cognitive tasks at the individual subject level [9], [66], [68], [90].

Despite this empirical success and theoretical merit, the simple characterization of decision-making encapsulated by EAMs results in somewhat restrictive applications [19]. In the context of visual tasks, EAMs do not typically provide a detailed specification of how raw visual stimuli are transformed into evidence and consequently do not explain the visual processing steps underlying decision-making. A number of recent modeling efforts have addressed this limitation by adapting convolutional neural network (CNN) models used in image classification tasks to the novel purpose of generating RTs or RT proxies [1], [24], [30], [45], [62], [80], [86], [87], a strategy we also pursue here. CNNs are useful in this regard since they are image-computable—they accept arbitrary images as input—and capture important characteristics of biological vision [43], [47], [95]. Early CNN layers exhibit spatially-organized units with local receptive fields, analogous to the retinotopic organization of the early mammalian visual pathway. The concatenation of multiple layers forms a hierarchy in which progressively more abstract features are extracted in deeper layers, in a way that is globally similar to the primate visual cortical hierarchy.

Here, we integrate CNN models for visual feature extraction with traditional EAMs of speeded decision-making tasks in a framework we call the visual accumulator model (VAM). As in the prior models referenced above, the VAM accepts raw (pixel-space) visual stimuli as inputs and generates RTs and choices as outputs. The key feature of the VAM that distinguishes it from prior models is that the CNN and EAM parameters are *jointly fitted* to the RT, choice, and visual stimulus data from individual participants in a unified Bayesian framework. Thus, both the visual representations learned by the CNN and the EAM parameters are directly constrained by behavioral data. In contrast, prior models first optimize the CNN to perform the behavioral task, then separately fit a minimal set of high-level CNN parameters [62] and/or the EAM parameters to behavioral data [1], [30], [87]. As we will show, fitting the CNN with human data—rather than optimizing the model to perform a task—has significant consequences for the representations learned by the model.

We leverage the VAM to explore how abstract, task-relevant information is extracted from raw sensory inputs, and we investigate how the behavioral phenomenon of congruency effects arises as a consequence of the representation geometry learned by the CNN. In doing so, our framework also addresses one of the main criticisms of deep neural network models of vision that are optimized to perform particular tasks (e.g., object identification): these models account for few results from psychology [3], [6], [20], [48], [50].

Modeling framework

We first describe the task—Lost in Migration (LIM), available as part of the Lumosity cognitive training platform—which will serve to ground the discussion of the model. LIM is a stylized version of the well-known arrows flanker task used in the study of cognitive control and visual attention (**Fig. 1A**; [82]). In LIM, participants are shown stimuli composed of several arrow-shaped birds. The task is to indicate the direction of the central bird (the target) using the arrow keys on the keyboard, ignoring the direction of the surrounding birds (the flankers). Participants engage with the task at home and are rewarded via a composite score that takes into account both speed and accuracy.

The stimuli used in LIM vary along several dimensions that are not present in the standard flanker task, implying potentially complicated stimulus-behavior dependencies that are well-suited to the VAM. First, both targets and flankers can be independently oriented left, right, up, or down, such that there are four possible response directions (the standard arrow flanker task allows for only left and right responses). Second, the layout of the targets and the flankers can appear in one of seven configurations: left/right/up/down 'V', horizontal/vertical line, and cross (**Fig. 1A**). Last, the target can be centered anywhere within the 640×480 pixel game window, subject to edge constraints.

As with other flanker tasks, a given trial is said to be congruent if targets and flankers are oriented in the same direction, and incongruent otherwise. A consistent observation from studies employing the flanker task and related conflict tasks are *congruency effects*: participants are slower and less accurate on incongruent trials [18], [78], [83]. Congruency effects are considered to index a specific aspect of cognitive control: they indicate the extent to which an individual can selectively attend to task-relevant information and ignore irrelevant information.

The visual accumulator is an image-computable EAM, composed of a CNN and EAM chained together (**Fig. 1B**). Minimally processed (pixel-space) visual stimuli are provided as inputs to the CNN. The outputs of the CNN correspond to the mean rates at which evidence is accumulated (the drift rates), one for each possible response (four in the case of LIM). The EAM then generates choices and RTs through a noisy evidence accumulation process. For each participant in the dataset, we fit one such model (CNN + EAM) jointly using that participant's visual stimuli, RTs, and choices. Building on prior work [12], [44], we developed an automatic differentiation variational inference (ADVI) algorithm that simultaneously optimizes the parameters of the CNN and learns the posterior distribution over the LBA parameters (Methods).

The particular EAM we adopt is the linear ballistic accumulator model (LBA [9]; **Fig. 1B**) since it can be applied to tasks with more than two possible responses, though the general VAM framework is compatible with other EAMs that have a closed-form or easily approximated likelihood. The LBA parameters fitted in the VAM are the decision threshold b, the non-decision time t_0, and a parameter A that controls the dispersion of the initial accumulator values (the drift rate means are fitted implicitly via the CNN). We used a seven layer CNN architecture (6 convolutional layers, 1 fully-connected layer) in the VAM (**Fig. 1B**). To speed up the training process, the first two convolutional layers were initialized with the parameters from a 16-layer VGG CNN trained to classify the ImageNet dataset [13], [79].

All of the code (*https://github.com/pauljaffe/vam*) and data (*https://doi.org/10.5281/zenodo.10775514*) used to train the VAMs and reproduce our results are publicly-available.

Results

Models capture human behavioral data

We fitted a separate VAM to LIM data (visual stimuli, RTs, choices) from each of 75 Lumosity users (participants). We selected participants who had practiced Lost in Migration extensively (≥ 25,000 trials) and used data from a later stage of practice to minimize learning effects (≥ each participant's 50th gameplay). These participants varied in age (23–87 years) and in their behavior (mean RT, accuracy, congruency effects), allowing us to examine how well our framework captured individual differences.

To assess the model fits, we compared the mean RT, accuracy, RT congruency effect (incongruent trial mean RT minus congruent trial mean RT), and accuracy congruency effect (congruent trial accuracy minus incongruent trial accuracy) of each model/participant pair on a set of holdout

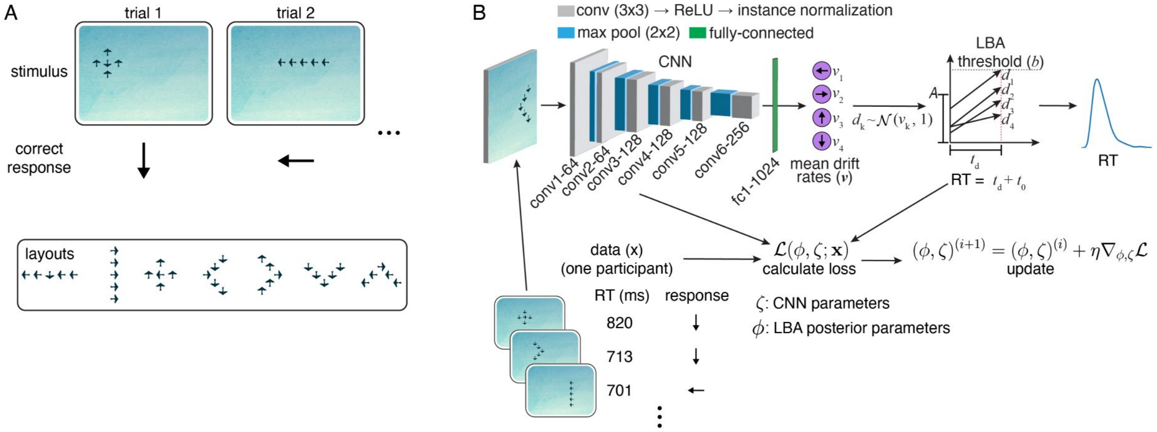


Figure 1

Task and model.

A) Top, Lost in Migration task. Bottom, the seven stimulus layouts (random target/flanker directions). **B)** VAM schematic. The numbers after the CNN layer names correspond to the number of channels used in that layer. See Methods for additional details.

stimuli, separate from those used to train the models (**Figs. 2** and **S1**). For each behavioral summary statistic, the responses of the fitted models were highly correlated with those of the participants (Pearson's $r > 0.75$), with slopes close to unity (**Fig. 2B–E**, statistics in figure legend). We found that the RT congruency effect could be attributed to a reduction in the mean target drift rate parameter on incongruent vs. congruent trials, while the accuracy congruency effect could be attributed to a higher mean flanker drift rate on incongruent trials relative to the non-target (other) mean drift rates on congruent trials (**Fig. 2F**). The latter observation follows from the fact that the drift rates on trial i are sampled from $N(v^{(i)}, 1)$, where the $v^{(i)}$ are the mean drift rates shown in **Fig. 2F**. Since the flanker drift rates on incongruent trials are higher (less negative) than the non-target drift rates on congruent trials, more errors result on incongruent trials, giving rise to the accuracy congruency effect.

We also examined whether the fitted models captured demographic effects, focusing on the well-established slowing of RTs that occurs with age [25], [56], [65]. Mean RTs began increasing around age 50, effects captured by the fitted models (**Fig. 2G**). Consistent with prior work in a variety of decision-making tasks [22], [65], [76], [81], we found that an age-dependent increase in the non-decision time (t_0) component of the response contributed to longer RTs in the models from older adults (**Fig. S2A**). In contrast to these prior studies, we did not observe increased response caution (measured as $b - A$) in the models from older adults (**Fig. S2B**). This discrepancy could be explained by differences in the task design or the fact that our participants had practiced the task substantially more than in typical studies [81]. We also observed an age-dependent reduction in the target drift rates and no age-dependence of the flanker drift rates (**Fig. S2C–D**), findings that have received mixed support in the literature [4], [22], [65], [76], [81].

One virtue of the VAM is that the model implicitly learns which stimulus properties influence behavior. As a simple demonstration of this, we investigated how two high-level visual features influenced participant behavior—stimulus layout and both horizontal/vertical stimulus position—and whether the fitted models captured these effects. We focused on RT effects, since no participants exhibited a significant effect of layout or horizontal/vertical position on accuracy ($p > 0.05$ for all stimulus features, chi-squared test), though we note that ceiling effects resulting from the high accuracy of the participants may have hindered our ability to detect these accuracy effects.

The majority of participants exhibited layout-dependent RT biases ($p < 0.05$ for 60/75 participants, one-way ANOVA; see examples in **Fig. 2H** and **Fig. S1B**). We quantified how well the models captured these effects by calculating the Pearson's r between model/participant mean RTs across each layout for the participants that exhibited significant layout-dependent RT modulation (**Fig. 2J**). The median Pearson's r was 0.67, demonstrating good correspondence between model/participant behavior. There was considerable heterogeneity in the particular layout RT biases exhibited by both the participants and models, though responses to trials with the vertical line layout were on average somewhat faster than the other layouts for both participants and models (**Figs. S1B** and **S3A**).

The majority of participants also exhibited both horizontal and vertical position-dependent RT biases (horizontal position: $p < 0.05$ for 72/75 participants, vertical position: $p < 0.05$ for 69/75 participants, one-way ANOVA; see examples in **Fig. 2I** and **Fig. S1C–D**). The fitted models captured these effects adequately: the median Pearson's r between model/participant mean RTs across stimulus position bins was 0.78 for horizontal position and 0.55 for vertical position (**Fig. 2J**). While there was substantial variability across both participants and models in the particular position-dependent RT biases they exhibited, most participants/models responded more slowly when the stimuli were positioned close to the horizontal edges and, to a lesser extent, the vertical edges of the task window (**Figs. S1C–D** and **S3B–C**).

We also examined whether the VAMs captured two other commonly used behavioral metrics that take into account additional information from the RT distribution: RT delta plots and conditional accuracy functions [35], [71], [88], [92], [93]. The participant RT delta plots showed that the congruency effect increased with longer RTs [61], a trend that was captured by the fitted VAMs (**Fig. S4A**). In contrast, we observed a mismatch between model and participant behavior for the conditional accuracy functions on incongruent trials (**Fig. S4B**). In particular, the participants tended to be less accurate on faster trials, while the models exhibited the opposite trend. This discrepancy may be explained by the lack of dynamic visual processing in the VAM, which has been proposed to account for the preponderance of errors on faster trials often observed in conflict tasks [92], [93]. In the remainder of our analyses, we focus on the mean congruency effect, leaving a full accounting of such dynamic error patterns for future modeling work.

In summary, the VAM captured individual differences in RTs, accuracy, and congruency effects, and the dependence of RTs on stimulus layout and position. To lay the groundwork for understanding how the learned visual representations of the models relate to these behavioral effects, we sought to characterize the general properties of these representations that enable proficient execution of the task.

Representations of task-relevant stimulus information

To perform LIM well, the models must learn representations that select the task-relevant information (target direction) and diminish the influence of irrelevant information (flanker direction, stimulus layout, and stimulus position). To investigate these representations, we presented the model with a holdout set of LIM stimuli (separate from the training stimuli) and analyzed the resulting unit activity in each CNN layer [29], [98] (**Fig. 3A**, Methods). The activations from a given layer l form an $N \times K_l$ matrix, where N is the number of stimuli in the holdout set (5000) and K_l is the number of active units in layer l (a variable fraction of units in each layer did not respond to any stimuli and were excluded from the activation matrix).

To characterize the emergence of target selectivity, we adopted analysis techniques from the neuroscience literature that aim to determine what stimulus properties are coded by population-level neural activity, analogous to the high-dimensional CNN activations studied here. Specifically, we quantified how well the target direction could be decoded from the activation matrix of a given layer with a linear support vector machine [31], [42], [74] (SVM; **Fig. 3B**). Note that only incongruent trials were used in these analyses, since the classifier could use the flanker direction to classify the target on congruent trials, artificially inflating performance. Decoding performance on holdout data for target direction was near chance (27%) in the first network layer and increased to nearly perfect decoding (≥ 97%) at layer 4, with a sharp increase between the second and third convolutional layers (**Fig. 3B**). The increase in target decoding accuracy from shallower to deeper network layers is generally consistent with neural recording studies in mammals and other neural networks studies that document more accurate decoding of abstract variables (e.g., object or category identity) in higher visual/cortical regions and deeper neural network layers [8], [53], [85].

We also investigated target selectivity at the single-unit level by quantifying the mutual information between each unit's activity and target direction [53], [85] (**Fig. 3C**). In contrast to the population-level decoding results, information for target direction exhibited a non-monotonic "hunchback" profile across the convolutional layers, then increased sharply in the final fully-connected layer. The observation that single-unit information for target direction decreased between the fourth and final convolutional layers indicates that the units become progressively less selective for particular target directions. Since population-level decoding remained high in these layers, this suggests a transition from representing target direction with specialized "target neurons" to a more distributed, ensemble-level code. Notably, a similar transition in coding properties takes place along the cortical hierarchy, with higher-order cortical regions exhibiting a

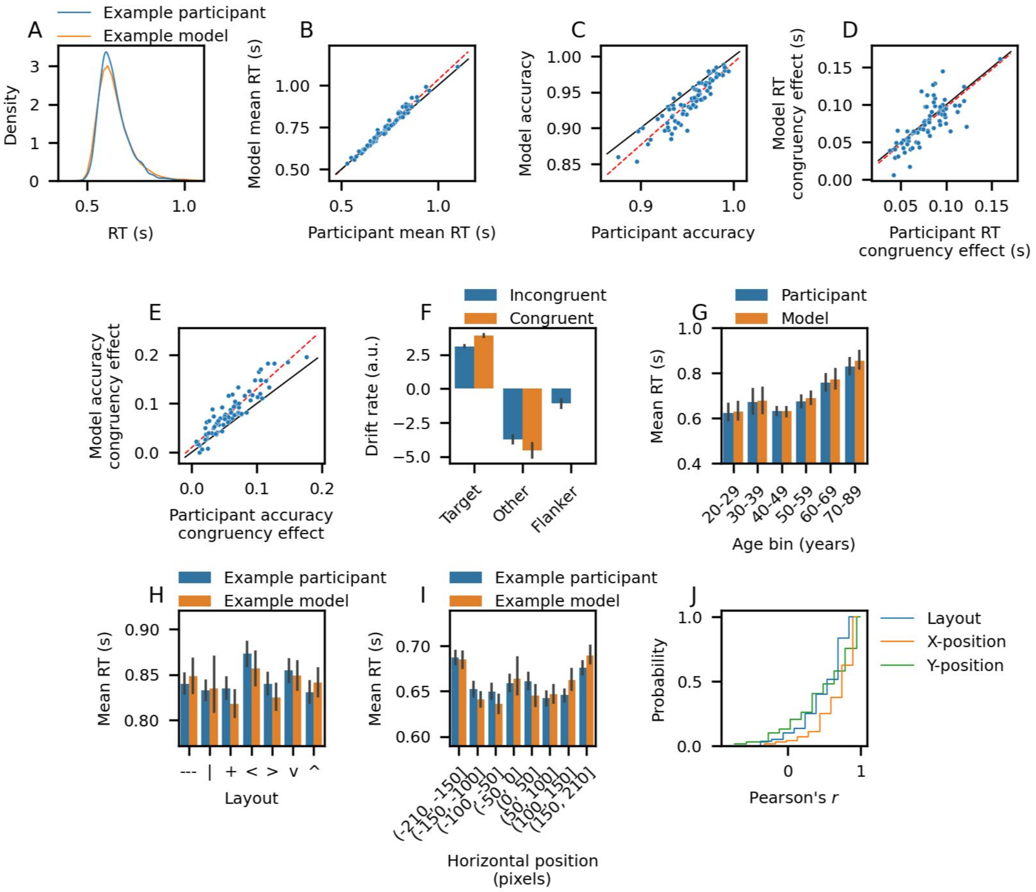


Figure 2

Comparison of model/participant behavior.

For panels B–E, each point is one model/participant (n = 75), black line: unity, red line: linear best-fit. **A)** Example model/participant RT distributions. **B)** Mean RT (Pearson’s r = 0.99, bootstrap 95% CI = (0.99, 0.99), best-fit slope = 1.07). **C)** Accuracy (r = 0.91, 95% CI = (0.87, 0.94), slope = 1.15). **D)** RT congruency effect (r = 0.77, 95% CI = (0.67, 0.86), slope = 1.01). **E)** Accuracy congruency effect (r = 0.92, 95% CI = (0.88, 0.94), slope = 1.20). **F)** Drift rates averaged across all trials and models. **G)** Mean RT vs. age averaged across models. **H)** Example model/participant mean RT vs. stimulus layout (Pearson’s r = 0.67). **I)** Example model/participant mean RT vs. horizontal stimulus position (negative values: left of center; Pearson’s r = 0.79). **J)** Empirical CDF of Pearson’s r between model/participant mean RTs across stimulus feature bins (only participants with significant RT modulation are shown; layout: n = 60 models/participants, x-position: n = 72, y-position: n = 69). Error bars in panels F–I are bootstrap 95% confidence intervals.

reduction in units with “pure” selectivity and a corresponding increase in units with mixed selectivity for multiple task or stimulus features [51], [52], [72]. The high proportion of units with mixed selectivity results in a high-dimensional representation in which all task-relevant information can easily be extracted with simple linear decoders [11], [57].

Consistent with these ideas, we found that the dimensionality of target representations as measured by the participation ratio (Methods) increased sharply in the last two convolutional layers(Conv5–Conv6), paralleling the reduction in single-unit information for target direction in these layers (**Fig. 3D**). The high-dimensional representation observed in these layers may enable the VAM to capture the rich stimulusbehavior dependencies present in the participant data. The reduction in dimensionality in the final fully-connected layer, and concomitant increase in single-unit information for target direction, may reflect a strong constraint to select the correct target direction imposed by the task. Notably, the hunchback-shaped profile we observed in the dimensionality of representations has also been observed along visual cortical regions of the rat ventral stream and in other neural network studies [2], [53].

To more explicitly characterize the degree of directional selectivity in each layer, we quantified the proportion of units that responded preferentially to one of the four target directions. We separated units according to the sign of modulation and degree of directional selectivity: “selective (+)” and “selective (-)” units were more (or less) active for one target direction relative to the other three; “complex” units exhibited significant modulation by target direction without a clear directional preference (Methods).

The proportion of units that were selective for a particular direction increased steadily from the second to the fourth convolutional layer, with most units exhibiting positive modulation (**Figs. 3E** & **S5**). Between the fourth and final convolutional layers, the proportion of units with positive target direction modulation decreased, while the proportion of units with negative and complex modulation increased.

In summary, a strong representation for target direction emerged in the middle convolutional layers, and was initially supported by a simple low-dimensional code, with information for each target direction concentrated in separate populations of directionally-tuned units. In later convolutional layers, target direction was supported by a more distributed, complex, and high-dimensional code, with weaker directional selectivity at the single-unit level.

Extraction and suppression of distracting stimulus information

How do the models’ visual representations enable target selectivity for stimuli that vary along several irrelevant dimensions? The mammalian visual system solves the analogous problem of visual object recognition through representations that become increasingly invariant to different object positions, sizes, and lighting conditions [15], [74]. To determine whether the models learned representations that share these characteristics, we initially assessed whether the model representations for target direction are indeed invariant (or more generally, tolerant), to variation in flanker direction, stimulus layout, and horizontal/vertical position. Alternatively, the models could have learned a coding scheme in which different representations are responsible for selecting the target in each distracter context, e.g., with units that respond to the conjunction of particular target direction and layout combinations.

To assess the degree of representation tolerance, we quantified how well a linear SVM trained to classify target direction in one distracter context generalized to a different distracter context, where the distracter context is defined by a particular type of task-irrelevant information [5], [74]. For example, to assess the degree of tolerance to flanker direction, we split trials into a training set where the flanker direction was fixed to a particular value (e.g. down), and a

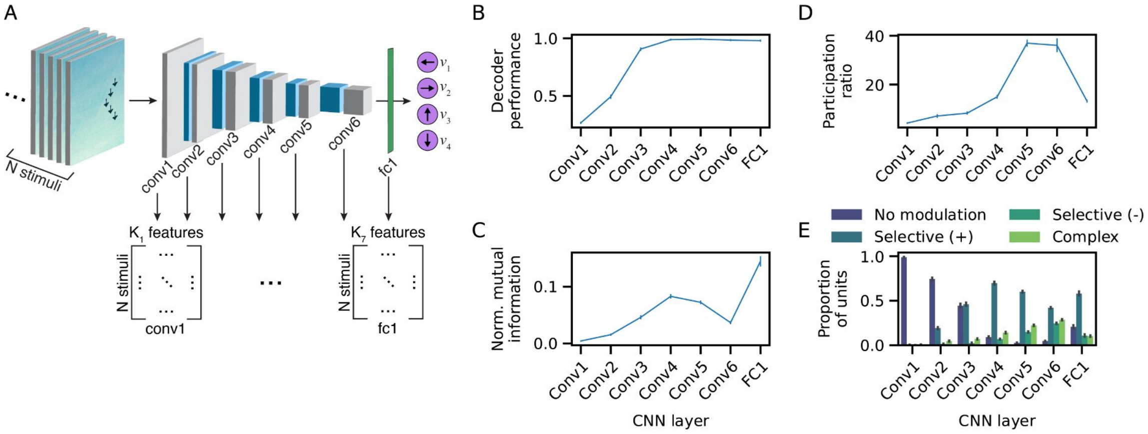


Figure 3

Neural representations of target direction.

A) Schematic of the CNN activations extracted from each network layer. Each layer yields a $N \times K_l$ activation matrix, where N is the number of stimuli and K_l is the number of active units (i.e., feature dimensions) in layer l. **B)** Decoding accuracy of stimulus target direction. **C)** Normalized mutual information for target direction conveyed by single units, averaged across units. Mutual information was normalized by the entropy of the target direction distribution (possible range = [0, 1]). **D)** Dimensionality of target representations as measured by the participation ratio of the target-centered activation covariance matrix. **E)** Proportion of units exhibiting selectivity for target direction. Panels B–E show the average across n = 75 models; error bars correspond to bootstrap 95% confidence intervals.

generalization set with all other flanker directions (e.g., flanker direction = left/right/up). The performance of the classifier on the generalization set measures the extent to which the target representations are tolerant (or invariant) to variability in a given type of distracting information.

For each type of distracting information, we found that generalization performance was initially near chance (25%) and increased steadily through the network layers (**Fig. 4A**). The lower generalization performance for flanker direction observed in the deepest network layers (~60%) can be attributed to the robust congruency effects exhibited by the models (**Fig. 2D–E**): for the vast majority of incongruent error trials, the model chose the flanker direction, implying that flanker direction has a strong impact on model representations.

The tolerance of target direction representations to variability in irrelevant stimulus features suggests that the irrelevant information was progressively suppressed from shallower to deeper network layers. To examine this explicitly, we quantified how well each irrelevant stimulus feature (flanker direction, stimulus layout, horizontal/vertical position) could be decoded from the activity in each layer using the same SVM classifier methodology as we did for target direction decoding. We found that decoding accuracy for flanker direction and stimulus layout exhibited a hunchback profile in which decoding performance started low in early layers, increased in intermediate layers, and decreased in later layers (**Fig. 4B**). The decoding accuracy for horizontal and vertical position followed a similar but shifted pattern: there was a steady increase in accuracy until the second-to-last layer, followed by a slight drop in the last layer. Partial (rather than complete) suppression of irrelevant stimulus features is expected given that these features all impact behavior (**Fig. 2**). Note that the increase in receptive field size that occurs from shallower to deeper layers is necessary but not *sufficient* for accurate decoding of the irrelevant stimulus features, and does not explain the reduction in decoding accuracy in later layers.

We also examined suppression at the single-unit level by quantifying the mutual information between each unit's activations and the values of each task-irrelevant stimulus feature [85] (**Fig. 4C**). The mutual information for a given stimulus feature was normalized by the entropy of the feature distribution to facilitate comparisons between the different stimulus features [53]. In agreement with the population-level decoding analyses, information for the irrelevant stimulus features exhibited a pronounced hunchback profile in the progression from shallower to deeper network layers. It is noteworthy that the suppression of irrelevant information is more pronounced at the single-unit level in that decoding accuracy remains relatively high in later layers, particularly for flanker direction. This parallels the findings discussed above for target direction, and again suggests a transition from a simple code with populations of units that are selective for particular stimulus features to a more distributed, ensemble-level code.

Orthogonality of task-relevant and irrelevant information predicts behavior

A noteworthy feature of visual attention is that the selectivity and tolerance identified above is not absolute: irrelevant information cannot be filtered out completely, as illustrated by the congruency effects observed in the flanker task and related conflict tasks. A common framework for understanding these behavioral phenomena posits that task-relevant and irrelevant information compete for control over response execution, and that congruency effects arise from incomplete suppression of irrelevant information [88], [93]. Motivated by these theories, we investigated whether the degree of suppression of irrelevant information in the trained models was correlated with congruency effects across the models we analyzed.

To this end, we operationalized suppression with two metrics of the model representations that we investigated above: the accuracy of decoders trained to classify flanker direction from the model representations and the mutual information for flanker direction conveyed by single units. We expected to observe a positive correlation between both of these metrics and both RT and accuracy congruency effects: higher decoder accuracy or mutual information for flanker direction corresponds to less suppression and therefore higher congruency effects. However, we did not observe a significant positive correlation between either decoding accuracy or mutual information and RT or accuracy congruency effects in any of the model layers, with the exception of a single significant positive correlation between flanker mutual information and accuracy congruency effects in layer Conv3 (**Fig. S6**).

Given this somewhat surprising negative result, we were motivated to consider an alternative account of congruency effects, one that takes into account the relative geometry of the task-irrelevant (flanker) *and* task-relevant (target) information. In particular, we considered the possibility that task-relevant and irrelevant information could be orthogonalized in the high-dimensional space of neural activity, such that task-relevant information is shielded from distracter interference. The general idea that neural representations for different types of information can be orthogonalized to prevent interference has received support from a number of neural recording studies [21], [37], [46], [58], [73].

To examine whether target and flanker representations are orthogonalized, we first defined target and flanker subspaces from the target direction and flanker direction classifiers used in the decoding analyses described above (Methods). The target direction classifier for a given network layer implicitly defines four decoding vectors, one for each target direction [5], [46]. We define the subspace of the K_l-dimensional feature space spanned by these four vectors as the target subspace; the flanker subspace is defined analogously.

To measure the orthogonality between target and flanker subspaces, we calculated the average of the cosine of the principal angles between the target and flanker subspaces, a metric we refer to as subspace alignment (Methods). Principal angles generalizes the idea of angles between lines or planes to arbitrary dimensions [36], [99]. The subspace alignment metric has a simple interpretation: it is equal to one if the subspaces are completely parallel and zero if they are completely orthogonal.

We first characterized how target/flanker subspace alignment develops across the layers of the trained models (**Fig. 5A**). Given that target direction decoding is poor in the first two layers (< 50%; **Fig. 3B**), the decoding vectors used to define the target subspace are not particularly meaningful, and we do not attempt to interpret the subspace alignment metric in these layers. Beginning at the third convolutional layer, when decoding accuracy for both targets and flankers is high (> 90%), we found that target and flanker subspaces are well-aligned. We interpret the high alignment as evidence that the model has learned a common representation for direction that is shared for both targets and flankers. In later layers, we found that target and flanker representations become increasingly orthogonal, consistent with the view that the processing in later layers acts to reduce interference between task-relevant and irrelevant information.

If orthogonalizing target and flanker representations reduces interference from the irrelevant (flanker) information, we should observe a positive correlation between subspace alignment and congruency effects across models, since greater alignment results in more interference. Consistent with this idea, we observed a significant positive correlation between subspace alignment and accuracy congruency effects across models in each layer beginning with the fourth convolutional layer (**Fig. 5B–C**; adjusted p-value < 0.05 for layers 3-7, permutation test, Bonferroni correction for 7 comparisons). In contrast, we did not observe a significant correlation between target/flanker subspace alignment and RT congruency effects in any network layer, suggesting a mechanistic dissociation between RT and accuracy congruency effects (**Fig. S7**).

Figure 4

Suppression of task-irrelevant information and tolerance in task-relevant representations.

A) Decoding accuracy of stimulus target direction in a new distracter context (generalization performance). Context was defined by the values of a given stimulus feature (flanker direction, layout, horizontal/vertical position). **B)** Decoding accuracy of irrelevant stimulus features. **C)** Normalized mutual information for irrelevant stimulus features conveyed by single units, averaged across units. For each stimulus feature, the mutual information was normalized by the entropy of the stimulus feature distribution (possible range = [0, 1]). All panels show the average across *n* = 75 models; error bars correspond to bootstrap 95% confidence intervals.

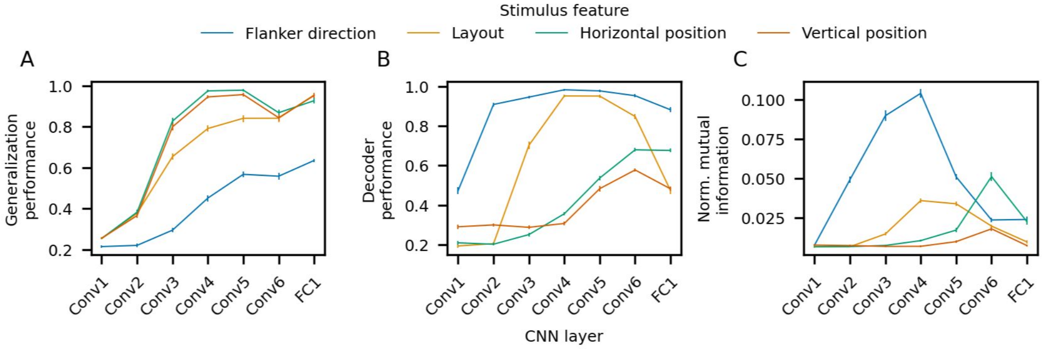
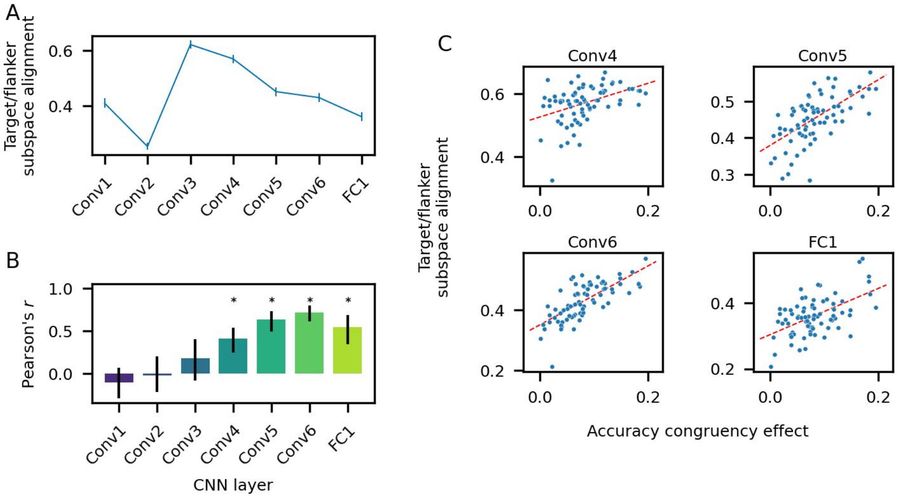


Figure 5

Orthogonality of target/flanker subspaces predicts accuracy congruency effects.

A) Target/flanker subspace alignment averaged across models. **B)** Pearson's correlation coefficient between target/flanker subspace alignment and accuracy congruency effect calculated across models. **C)** Target/flanker subspace alignment vs. accuracy congruency effect for layers Conv4–FC1. Each point corresponds to one model; the red line is the linear best-fit. For all panels, *n* = 75 models. Error bars in panels A–B correspond to bootstrap 95% confidence intervals. Asterisks in panel B indicate a significant Pearson's *r* (adjusted *p*-value < 0.05, permutation test with *n* = 1000 shuffles, Bonferroni correction for 7 comparisons).



Representation geometry of task-optimized models

Researchers who use neural network models to study neural representations typically optimize the model to perform a task, rather than fit the model to behavioral data from the task as we do here [51], [91], [96], [97], cf. [14], [33], [84]. This raises the possibility that these two training paradigms induce different representations in the models. To investigate this, we trained CNNs to perform LIM (i.e., output the direction of the target) by minimizing the standard cross-entropy loss function used in image classification tasks, where the training labels are given by the true direction of the target bird in each stimulus. The task-optimized CNNs were identical to those used in the VAMs, except that the outputs of the last layer were converted to softmax-scored probabilities for each direction rather than drift rates. Otherwise, all aspects of the optimization algorithm, CNN architecture, initialization, and training data for the task-optimized models were the same as those used to train the VAMs. We trained one task-optimized model for each VAM using stimulus data from the same participants (n = 75 task-optimized models).

We first compared the behavioral outputs of the task-optimized models and the VAMs. We found that the task-optimized models did not exhibit an accuracy congruency effect (**Fig. 6A**). Thus, simply training the model to perform the task is not sufficient to reproduce a behavioral phenomenon widely-observed in conflict tasks. This challenges a core (but often implicit) assumption of the task-optimized training paradigm, namely that training a model to do a task well will result in model representations that are similar to those employed by humans. Indeed, for a number of visual tasks, the representations and behavior of task-optimized CNNs has been observed to differ considerably from those of humans [17], [32], [34], [63], [75].

Since the task-optimized models do not generate RTs, it is not possible to directly measure RT congruency effects in these models without making additional assumptions about how the CNN's classification decisions relate to RTs. However, as a coarse proxy for RT, we can examine the confidence of the CNN's decisions, defined as the softmax-scored logit (probability) of the most probable direction in the final CNN layer. This choice of RT proxy is motivated by some prior studies that have combined CNNs with EAMs [1], [30], [87]. These studies explicitly or implicitly derive a measure of decision confidence from the activity of the last CNN layer. The confidence measure is then mapped to the EAM drift rates, such that greater decision confidence generally corresponds to higher drift rates (and therefore shorter RTs).

We calculated the average confidence of each task-optimized CNN separately for congruent vs. incongruent trials. On average, the task-optimized models showed higher confidence on congruent vs. incongruent trials (W = 21.0, p < 1e-3, Wilcoxon signed-rank test; Cohen's d = 0.99; n = 75 models). These analyses therefore provide some evidence that task-optimized CNNs have the capacity to exhibit congruency effects, though an explicit comparison of the magnitude of these effects with human data requires additional modeling assumptions (e.g., fitting a separate EAM).

Above, we showed that VAMs with greater orthogonalization of target and flanker information exhibit smaller accuracy congruency effects (**Fig. 5B**), providing evidence that the relative geometry of task-relevant and irrelevant representations is a critical determinant of the degree of flanker interference at the behavioral level. Given that the task-optimized models do not exhibit an accuracy congruency effect, we therefore expect that these models would exhibit a higher degree of orthogonalization of target and flanker information. Consistent with this idea, we found that the task-optimized models had lower target/flanker subspace alignment (i.e., higher orthogonalization) for all network layers beginning with the third convolutional layer (**Fig. 6B**).

A direct consequence of the training paradigms used to train the VAMs is that these models are encouraged to capture dependencies between the stimulus features and behavior (RTs and accuracy), while the task-optimized models are not. As a result, the VAMs may learn more complex representations of the stimuli, since a variety of stimulus features—layout, stimulus position, flanker direction—influence behavior (**Fig. 2**). To investigate this possibility, we compared the dimensionality of target representations between the VAMs and task-optimized models using the participation ratio metric discussed above.

We found that the dimensionality of the VAM and task-optimized model representations was nearly identical for the first four convolutional layers (**Fig. 6C**). In contrast, for the final two convolutional layers, the VAMs exhibited a substantially more pronounced expansion of dimensionality than the task-optimized models. In the final fully-connected layer, dimensionality decreased sharply for both types of models, and was somewhat higher for the VAMs.

The increased dimensionality of VAMs' target represenations in later network layers is consistent with the view that these models must learn more complex representations of the stimuli in order to successfully capture stimulus-behavior dependencies. It is also noteworthy that the most striking difference between the dimensionality of the VAMs and task-optimized models occurs during the latter part (layers Conv5–Conv6) of the expansion phase of the hunchback-shaped dimensionality profile discussed above and observed in prior work [2], [53]. In these layers, single-unit information for target and flanker direction—the primary task features—decreases, while population-level decoding of these features remains high (**Figs. 3B–C** & **4B–C**). As discussed above, this dissociation implies a transition from a simple representation of target/flanker direction with separate populations of directionally-tuned units to a more complex and distributed code.

To determine whether the task-optimized models exhibited this change in coding properties, we quantified the single-unit information and population-level decoding accuracy for target/flanker direction in these models. Relative to the VAMs, the task-optimized models had substantially higher single-unit information for both target and flanker direction in layers Conv5–Conv6 (**Fig. 6D**). The task-optimized models also showed marginally more or roughly equivalent single-unit information for stimulus position in these layers relative to the VAMs, but had less single-unit information for stimulus layout (**Fig. S8A**).

In contrast, population-level decoding accuracy for target/flanker direction and stimulus position was similar between the task-optimized models and VAMs in layers Conv5–Conv6, though decoding accuracy for stimulus layout was notably lower for the task-optimized models (**Fig. 6E** & **S8B**). These results suggest that the task-optimized models maintained a simpler code for target/flanker direction in the later convolutional layers relative to the VAMs, primarily relying on separate populations of directionally-tuned units. Consistent with this idea, the task-optimized models had a higher proportion of simple selective (+) directionally-tuned units and a lower proportion of complex units in layers Conv5-6 relative to the VAMs (**Fig. 6F**).

Discussion

The dominant models of decision-making, while providing a good quantitative description of psychophysical data, do not incorporate biologically plausible models of the perceptual processes that are essential for many behaviors [19]. On the other hand, neural network models of vision, while capturing core properties of the primate visual system, do not account for many results from behavioral experiments [3], [6], [20], [48], [50]. The VAM addresses both of these limitations by integrating neural network models for visual processing with traditional decision-making models in a unified probabilistic framework that can be fitted to visual stimuli and RT data. Leveraging large-scale data from a task with rich visual stimuli, we demonstrate that our framework captures complex dependencies between stimulus features and behavior at the level of individual participants. We also illustrate how congruency effects—a core behavioral phenomenon observed in conflict tasks—can be explained in terms of the visual representations of the model. Finally, we document several key differences between the representations learned by models fitted to human behavioral data (the VAMs) and those learned by models trained only to do the task.

Processing phases underlying the transformation of sensory information

To perform the task and capture the behavioral data, the VAMs learned representations that—like the mammalian cortex—extract task-relevant information from raw visual inputs. Our analyses of these representations revealed several discrete processing phases that are fruitfully discussed in relation to the changes in target representation dimensionality we observed. Across the layers of the VAM's CNN, target dimensionality exhibited a prominent hunchback shape profile, corresponding to an initial protracted phase of dimensionality expansion followed by abrupt dimensionality compression in the final network layers. An analogous expansion and subsequent compression of object representation dimensionality has been documented in CNNs trained to classify images and along the rat visual-cortical hierarchy [2], [53], with some notable differences from our work that we highlight below.

We speculate that the initial phase of dimensionality expansion can be explained in part by the pruning of low-level stimulus features (e.g., contrast and luminosity) that are correlated across stimuli in the dataset [2]. These correlations are particularly strong in our dataset since we used the same background image for each stimulus, resulting in low-dimensional activity in the initial network layers. Conceivably, the initial increase in target representation dimensionality (i.e., layers Conv1–Conv4) results from the removal of these correlations, analogous to a whitening transformation [2].

In the early and middle convolutional layers (Conv1–Conv4), we found that the population-level decoding and single-unit information for both task-relevant (target direction) and distracting information (flanker direction, stimulus layout, and position) increased, occurring in parallel with the subtle expansion of dimensionality we observed in these layers. These trends are broadly consistent with observations that the mammalian visual cortex encodes increasingly abstract stimulus properties (e.g., object identity) in downstream brain regions [26], [31], [74]. The fact that information for distracting stimulus features increased in these layers is especially noteworthy, given that this information—the flanker direction, most prominently—impairs performance on the task [18], [82]. While it is tempting to speculate that the VAMs learned to extract distracter information because they were fitted to behavioral data, where the impairments in task performance manifest, the task-optimized models also exhibited increased distracter information in these layers. This suggests an alternative explanation in which the model first extracts more granular, superordinate stimulus representations that group together stimulus features with similar statistics. For example, the initial increase in encoding of target and flanker direction may reflect the representation of a unitary "directional signal" that facilitates the eventual separation of target and flanker direction representations that occurs in later layers. This idea is consistent with our observation that target/flanker subspaces are highly aligned in the middle convolutional layers and become progressively more orthogonal in later network layers.

In the last convolutional layers (Conv5–Conv6), we observed a large expansion in target representation dimensionality, which was notably less pronounced in the task-optimized models compared to the VAMs. We speculate that the increase in dimensionality may be partly due to an increase in mixed selectivity for multiple stimulus features at the single-unit level, similar to the single-unit coding properties observed in primate PFC [72]. In general agreement with this idea, we observed a reduction in the proportion of units with selectivity for particular target directions in these layers, and a concomitant increase in units that were modulated by multiple target directions. Relative to the VAMs, the task-optimized models showed a higher proportion of

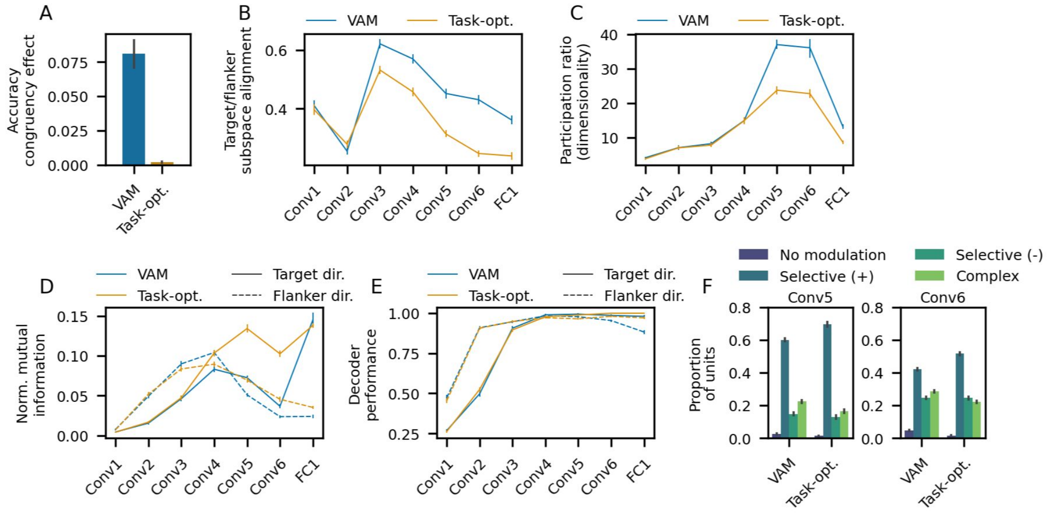


Figure 6

Comparison of VAMs and task-optimized models.

A) Accuracy congruency effect. **B)** Target/flanker subspace alignment. **C)** Dimensionality of target representations, as measured by the participation ratio of the target-centered activation covariance matrix. **D)** Normalized mutual information for target/flanker direction conveyed by single units, averaged across units. Mutual information was normalized by the entropy of the target/flanker direction distribution (possible range = [0, 1]). **E)** Decoding accuracy of target/flanker direction. **F)** Proportion of units exhibiting selectivity for target direction in layers Conv5–Conv6. All panels show the average across n = 75 task-optimized models and n = 75 VAMs; error bars correspond to bootstrap 95% confidence intervals. The VAM data shown in panels A–F is the same as that shown in **Figs. 2E**, **5A**, **3D**, **3C**/**4C**, **3B**/**4B**, and **3E**, respectively.

directionally-tuned units in these layers. A notable advantage of high-dimensional neural codes composed of units with mixed selectivity is that many different input-output relationships can be implemented with simple decoders [11], [57], [72]. Thus, the higher dimensionality and more complex target direction coding observed in the VAMs relative to the task-optimized models may reflect the fact that the VAMs are trained to capture a potentially large number of dependencies between stimulus attributes and behavior, while the task-optimized models need only extract the target direction. Given the impressive capacity of primates to learn complex tasks and arbitrary stimulus behavior relationships, and the abundance of high-dimensional representations across the cortex, models trained to capture the richness of stimulus-behavior dependencies—such as the VAM—may result in better models of cortical processing than optimizing models to perform tasks.

Congruency effects emerge from the relative geometry of task-relevant vs. distracting sensory representations

One of the key strengths of our modeling framework is that neural representations of stimulus features can be directly related to behavioral outputs. We focused in particular on congruency effects since they are ubiquitously observed in conflict tasks and highlight an inherent limitation of selective attention, namely that humans cannot completely filter out distracting information. Congruency effects are often interpreted in the context of a "dual-process" model in which automatic or impulsive processing arising from distracting stimulus information and more controlled or deliberate processing of task-relevant information compete for control over behavior [70], [88], [93]. According to this model, congruency effects result from the incomplete suppression of incorrect response activation in the automatic pathway. Building on this framework, a variety of models have been proposed that successfully capture congruency effects and other behavioral phenomena observed in conflict tasks [10], [88], [92].

Our work differs from these prior modeling efforts in two key ways. First, we did not attempt to "build in" a predetermined implementation of congruency effects. Rather, we fitted a relatively unstructured neural network model to human flanker task data and examined how congruency effects emerged [33]. Second, we explicitly modeled how task-relevant (and task-irrelevant) representations are extracted from raw visual stimuli with a biologically-plausible model of the mammalian visual system (a CNN). These features of the VAM allowed us to explore a space of possible explanations for congruency effects that are not readily investigated with other models of conflict tasks. Each CNN layer is composed of many simple neuron-like processing units, enabling a population-level description of task-relevant and irrelevant representations. Prior connectionist models of conflict tasks are also implemented as networks of interacting processing units [10], but are much reduced in scale relative to the VAM, precluding a characterization of stimulus feature representations in terms of their high-dimensional geometry as we pursue here. Another relevant attribute of CNNs is that the successive convolution and pooling operations in each layer extract increasingly abstract, task-related representations from raw visual inputs, a hallmark of the mammalian ventral visual system. This enabled a rich characterization of the intermediate visual representations that sequentially transform raw visual stimuli to behavioral outputs.

Leveraging these features of the VAM framework, we investigated potential explanations for congruency effects in terms of the population-level activity and representation geometry in each network layer. Motivated by the dual-process model of conflict tasks mentioned above, we initially sought evidence for the suppression of task-irrelevant information, as operationalized by population-level decoding accuracy and single-unit information for flanker direction. While we did find evidence of greater suppression in later vs. middle convolutional layers for both of these metrics, variability in the magnitude of this suppression across models was not related to variability in congruency effects.

Instead, we found that the relative geometry of target and flanker representations was a critical determinant of congruency effects. In particular, models with smaller accuracy congruency effects had more orthogonal (or less aligned) representations of targets and flankers in later network layers, proximal to behavioral outputs. This finding is consistent with the idea that orthogonalization of task-relevant/irrelevant representations shields the task-relevant information from distracters, and parallels findings from neural recording studies in both humans and animals that representations for different sources of information can be orthogonalized in order to prevent interference [21], [37], [46], [58], [73], [94].

To elaborate on this idea, in models with more aligned (less orthogonal) target/flanker representations in intermediate layers, the task-relevant (target) subspace is effectively "contaminated" by task-irrelevant (flanker) information. In the absence of any corrective process, the task-relevant subspace propagates a mixture of both correct (target) and incorrect (flanker) direction signals forward through the network. At the final readout layer of the network, assuming that the target subspace is well-aligned with the readout weights [59], flanker signals within the target subspace then contribute to the drift rate for the (incorrect) flanker direction. This increases the probability of an error, such that models with more aligned target/flanker representations exhibit larger accuracy congruency effects. A corollary of this proposition is that the absolute amount of flanker information in the network—equivalently, the degree to which flanker information has been suppressed—is not necessarily predictive of congruency effects. This may account for our observations that the measures of suppression we considered are not correlated with accuracy congruency effects across models, and that representations for flanker direction do not appear to be strongly suppressed even in later network layers, as evidenced by high decoding accuracy for flanker direction in both the VAMs and task-optimized models.

The role of dynamics in conflict tasks

One apparent limitation of the VAM as presented here is that it does not have visual processing dynamics, which seem to be required to explain some observations from the flanker task and related conflict tasks. For example, the RTs on incongruent error trials are typically faster than error RTs for congruent trials and RTs for correct trials [93], an effect that we confirm is also present in the flanker task variant studied here. This observation can be explained by a "shrinking attention spotlight" in which response activation from flankers starts high and diminishes over time, resulting in a higher proportion of errors for faster RTs [92]. We speculate that our models were unable to capture these particular error patterns because the visual processing module (CNN) we used does not have any dynamics (e.g., recurrence) that could instantiate such time-varying attention and resultant time-varying drift rates. However, it is not difficult to imagine how the orthogonalization mechanism described above, which explains variability in accuracy congruency effects *across* individuals, could act in concert with other dynamic processes that explain variability in congruency effects *within* individuals (e.g., as a function of RT). In general, any process that dynamically gates the influence of irrelevant sensory information on behavioral outputs could accomplish this, for example ramping inhibition of incorrect response activation [93], a shrinking attention spotlight [92], or dynamics in neural population-level geometry [37]. To pursue these ideas, future work may aim to incorporate dynamics into the visual component and decision component of the VAM with recurrent CNNs [24], [55] and the task-DyVA model [33], respectively.

Conclusion

The VAM is a probabilistic model of psychophysical data that captures how raw sensory inputs are transformed into the abstract representations that guide decisions. Raw (pixel-space) visual stimuli are processed by a biologically-plausible neural network model of vision that outputs the parameters of a traditional decision-making model. Each VAM is fitted to data from a single participant, a feature that allowed us to study how individual differences in behavior emerge from differences in the “brains” of the models. To this end, we found that models with smaller congruency effects had more orthogonal representations for task-relevant and irrelevant information. While we chose to use a CNN to model visual processing, we note that the VAM is not limited to this choice: other sensory encoding models, such as those based on transformer architectures [16], can be readily swapped in to replace the CNN with minimal changes to the underlying VAM implementation. Similarly, the LBA decision-making model we employed is easily replaced with other decision-making models that have a closed-form or easily approximated likelihood, such as the diffusion-decision model and leaky competing accumulator model [49], [54], [68], [90]. In this way, the VAM provides a general probabilistic framework for jointly fitting interpretable decision-making models and expressive neural network architectures of sensory processing with psychophysical data.

Methods

Datasets

We used deidentified Lost in Migration gameplay data from 75 Lumosity users (participants) to train the models included in this paper. The mean (SD) age of this sample at the time of signup was 56.4 (18.6) years; 46.7% identified as female, 48.0% identified as male, and 5.3% did not report their gender. All analysis and modeling was done retrospectively on preexisting Lumosity data that were collected during the normal use of the Lumosity program (at home). All participants consented to the use and disclosure of their deidentified Lumosity data for any purpose as laid out in the Lumosity Privacy Policy (*www.lumosity.com/legal/privacy_policy*). No statistical methods were used to predetermine sample sizes, though our sample sizes are comparable to or larger than those used in related modeling work [9], [92].

The included participants were selected from a larger pool that met certain inclusion criteria. The selection from this larger pool was done at random, with the following exception: we included all participants under the age of 40 (n = 19) to ensure adequate representation of younger participants. The criteria used to define the larger participant pool were as follows: we required that participants had signed up as Lumosity users between June 28, 2015 and June 30, 2020 (inclusive); that they were between the ages of 18 and 89 at the time of signup (inclusive); that their country of origin was Australia, Canada, New Zealand, or the United States; that their preferred language was English; and that they were not employees of Lumos Labs, Inc. Accounts created for research purposes were also excluded. We also required that all of a given participant’s Lost in Migration gameplays were done on the web (as opposed to mobile) platform. Finally, we required that participants had at least 200 Lost in Migration gameplays and at least 25,000 trials starting with their 50th gameplay. One additional participant with near chance accuracy (33%) on Lost in Migration was excluded from the larger participant pool.

Trials with very short and very long RTs were excluded and did not count toward the 25,000 trial minimum. Specifically, we excluded trials less than or equal to 250ms and trials classified as outliers from a criterion based on the median absolute deviation from the median (the MAD): trials with an absolute deviation from the median RT of more than 10 times a given participant’s MAD were excluded.

The final datasets from each participant used for model training consist of the first 25,000 non-outlier trials starting with their 50th gameplay. Data from the first 49 gameplays were excluded to reduce learning-related variability. Approximately 50% of trials were congruent (vs. incongruent).

Modeling framework

Linear ballistic accumulator (LBA) model

The decision-making component of our modeling framework is the LBA model [9], tailored to Lost in Migration. Evidence for each of the four possible response directions is accumulated linearly and independently at a response-specific drift rate d_k. Commitment to a decision occurs when one of the accumulators reaches a fixed threshold parameter b. The RT on a given trial is the duration of this evidence accumulation process plus a constant non-decision time parameter t_0 that includes both sensory processing and motor execution. Response variability comes from two sources.

First, on each trial, the drift rate d_k for each response $k \in \{1, 2, 3, 4\}$ is sampled independently from a Gaussian distribution with mean vk and a common SD s. We fix s to 1 for all models to ensure that the LBA parameters are identifiable [12], [27]. Second, the initial evidence for each accumulator is sampled independently from a uniform distribution on the interval $[0, A]$, where A is a model parameter.

Visual accumulator model (VAM): overview

The VAM is a generalization of the LBA model in which the drift rate means $v^{(i)} = \{v_1^{(i)}, v_2^{(i)}, v_3^{(i)}, v_4^{(i)}\}$ on trial i depend on the stimuli $s^{(i)}$ through a CNN. For a dataset with N trials, the task data are $\mathbf{x}$ = {response times $\mathbf{t}$, choices $\mathbf{c}$, stimuli $\mathbf{s}$}, where $\mathbf{t} = \{t^{(i)}\}_{i=1}^{N}$, $\mathbf{c} = \{c^{(i)}\}_{i=1}^{N}$, and $\mathbf{s} = \{s^{(i)}\}_{i=1}^{N}$. We adopt a Bayesian framework and model the joint density of the task data $\mathbf{x}$ and LBA parameters $\boldsymbol{\theta} = \{b, A, t_0\}$ as:

$$p(\mathbf{x}, \boldsymbol{\theta}) = \prod_{i=1}^{N} p(t^{(i)}, c^{(i)} | v^{(i)}, b, A, t_0) p(b, A, t_0), \tag{1}$$

$$v^{(i)} = \mathrm{CNN}_{\boldsymbol{\zeta}}(s^{(i)}), \tag{2}$$

where the drift rate means $v^{(i)}$ depend on a CNN with parameters $\boldsymbol{\zeta}$ (described below). The factor on the left of Equation (1) is the likelihood of the LBA model, which has a closed-form solution [9]. The factor on the right is the prior distribution over the LBA parameters, specified as a standard multivariate Gaussian $p(\boldsymbol{\theta}) \sim N(\mathbf{0}, \mathbf{I})$. We express the joint density more compactly as:

$$p(\mathbf{x}, \boldsymbol{\theta}; \boldsymbol{\zeta}) = p(\mathbf{t}, \mathbf{c} | \mathrm{CNN}_{\boldsymbol{\zeta}}(\mathbf{s}), b, A, t_0) p(b, A, t_0). \tag{3}$$

To fit the VAM, i.e., learn the posterior distribution over the LBA parameters $p(\boldsymbol{\theta} | \mathbf{x})$ and simultaneously optimize the CNN parameters $\boldsymbol{\zeta}$, we apply automatic differentiation variational inference (ADVI) [44]. Rather than attempting to sample from the true posterior directly as in Markov chain Monte Carlo (MCMC) methods, variational inference introduces an approximate posterior density $q(\boldsymbol{\theta}; \boldsymbol{\varphi})$ and minimizes the Kullback-Leibler (KL) divergence from $p(\boldsymbol{\theta} | \mathbf{x})$ to $q(\boldsymbol{\theta}; \boldsymbol{\varphi})$ by optimizing $\boldsymbol{\varphi}$. Here, we specify the approximate posterior $q(\boldsymbol{\theta}; \boldsymbol{\varphi})$ as a multivariate Gaussian with mean $\boldsymbol{\mu}$ and unconstrained covariance matrix Σ(thus $\boldsymbol{\varphi} = \{\boldsymbol{\mu}, \Sigma\}$).

We use variational inference since it scales well to large datasets and can handle complicated models (such as the VAM), in contrast to MCMC methods. Leveraging automatic differentiation software (JAX [7]), we automate the calculation of the derivatives of the variational objective (described below) with respect to both the CNN parameters $\boldsymbol{\zeta}$ and variational parameters $\boldsymbol{\varphi}$.

Variable transformations

Note that the LBA parameters $\boldsymbol{\theta} = \{b, A, t_0\}$ are restricted to be nonnegative, while $\boldsymbol{\mu} \in \mathbb{R}^3$ and $\sum$ is restricted to be positive semidefinite. However, to optimize the variational objective (defined below), we require that $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$ have the same support [44]. To achieve this, we transform $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$ to both have support on the real line. Specifically, we reparameterize $\sum$ using the Cholesky decomposition as $\sum = \mathbf{LL^T}$, where $\mathbf{L}$ is a lower-triangular matrix with entries in $\mathbb{R}$ (so that now $\boldsymbol{\varphi} = \{\boldsymbol{\mu}, \mathbf{L}\}$). For the LBA parameters, in addition to mapping them to the real line, we must also enforce the constraint that the threshold b is always greater than the parameter A controlling the range of the initial evidence distribution. We enforce this by defining $\tilde{b} = b - A$ and taking the log of $\tilde{b}$, A, and t_0 to map them to the real line [12]. We define the transformed parameters as $\boldsymbol{\theta}^* = \{b^*, A^*, t^*_0\} = \{\log (b — A), \log A, \log t_0\}$ and the transformation that maps $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$ as T.

VAM objective function

To fit the VAM, we maximize the evidence lower bound (ELBO):

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\zeta}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\theta})}[\log p(\mathbf{x}, \boldsymbol{\theta}; \boldsymbol{\zeta}) - \log q(\boldsymbol{\theta}; \boldsymbol{\phi})].$$

This is the standard ELBO optimized in variational inference, with an additional dependence on the CNN parameters $\boldsymbol{\zeta}$. The ELBO is a lower bound on the marginal likelihood of the data $p(\mathbf{x})$, and maximizing the ELBO is equivalent to minimizing the KL divergence from $p(\boldsymbol{\theta}|\mathbf{x})$ to $q(\boldsymbol{\theta}; \boldsymbol{\varphi})$. Writing the ELBO in terms of the transformed variables $\boldsymbol{\theta}^*$, we have:

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\zeta}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\theta}^*)}[\log p(\mathbf{x}, T^{-1}(\boldsymbol{\theta}^*); \boldsymbol{\zeta}) + \log|\det J_{T^{-1}}(\boldsymbol{\theta}^*)| - \log q(\boldsymbol{\theta}^*; \boldsymbol{\phi})]. \tag{4}$$

The term log $|\det J_{T^{-1}}(\boldsymbol{\theta}^*)|$ is a Jacobian adjustment for the transformation T^{-1} that maps $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}$, required to ensure that the transformed density integrates to one. For the transformation T^{-1}, with $\boldsymbol{\theta}^*$ vectorized as $[b^*, A^*, t^*_0]$, the Jacobian is given by:

$$J_{T^{-1}}(\boldsymbol{\theta}^*) = \begin{bmatrix} b^* & A^* & 0 \\ 0 & A^* & 0 \\ 0 & 0 & t_0^* \end{bmatrix}.$$

Taking the log of the absolute value of the determinant of $J_{T^{-1}}(\boldsymbol{\theta}^*)$ gives the required adjustment:

$$\log|\det J_{T^{-1}}(\boldsymbol{\theta}^*)| = \log|b^* A^* t_0^*|. \tag{5}$$

We will apply stochastic gradient ascent to maximize the ELBO objective function given by Equation (4), using Monte Carlo (MC) estimates of the expectation and AD to calculate gradients with respect to $\boldsymbol{\varphi}$ and $\boldsymbol{\zeta}$. However, the gradients of the ELBO with respect to $\boldsymbol{\varphi}$ cannot be calculated directly by AD, since the expectation in Equation (4) is taken with respect to $q(\boldsymbol{\theta}^*; \boldsymbol{\varphi})$, which depends on $\boldsymbol{\varphi}$. We work around this by applying the *reparameterization trick* [38], [69], which expresses $\boldsymbol{\theta}^*$ as a differentiable transformation of an auxiliary noise variable $\boldsymbol{\varepsilon} \sim p(\boldsymbol{\varepsilon})$,where $p(\boldsymbol{\varepsilon})$does not depend on $\boldsymbol{\varphi}$. In particular, we set $p(\boldsymbol{\varepsilon})$to be a standard multivariate Gaussian, $p(\boldsymbol{\varepsilon}) = N(\boldsymbol{\varepsilon}; \mathbf{0}, \mathbf{I})$, and define the transformation $\widetilde{\theta}^* = \mathbf{L}\epsilon + \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ and $\mathbf{L}$ are the mean and covariance parameters of the Gaussian approximating density $q(\boldsymbol{\theta}^*; \boldsymbol{\varphi})$.

Now we estimate the expectation in Equation (4) with MC samples from $p(\boldsymbol{\varepsilon})$. Since the expectation no longer depends on $\boldsymbol{\varphi}$, we can use AD directly on these MC estimates to obtain unbiased gradients of the ELBO. We estimate the ELBO for each data point using independent MC samples, since this reduces the variance of the gradients relative to using the same set of MC samples for a given batch of data [40]. Thus the ELBO objective for the *i*th data point is estimated by:

$$\begin{gathered}\widetilde{\mathcal{L}}^{(i)}(\boldsymbol{\phi}, \boldsymbol{\zeta}) = \frac{1}{L}\sum_{l=1}^{L} \log p(x^{(i)}, T^{-1}(\boldsymbol{\theta}^{*(i,l)}); \boldsymbol{\zeta}) + \log|\det J_{T^{-1}}(\boldsymbol{\theta}^{*(i,l)})| - \log q(\boldsymbol{\theta}^{*(i,l)}; \boldsymbol{\phi}), \\ \text{where} \quad \boldsymbol{\theta}^{*(i,l)} = \mathbf{L}\boldsymbol{\epsilon}^{(i,l)} + \boldsymbol{\mu}, \quad \boldsymbol{\epsilon}^{(i,l)} \sim p(\boldsymbol{\epsilon}), \quad \text{and} \quad x^{(i)} = \{t^{(i)}, c^{(i)}, s^{(i)}\}.\end{gathered} \tag{6}$$

VAM inference algorithm and other training details

The VAM training/inference algorithm is summarized in Algorithm 1. The size of the training set was 16,250 samples (65% of the total dataset) for each model. An additional validation set of 3,750 samples (15%) was used to monitor model training. The remaining 5,000 samples (20%) were used to evaluate model performance (the holdout set). We set the batch size *M* to 256 and the number of MC samples used to estimate the ELBO *L* to 10. We used the Adam optimizer [39] with the following hyperparameters for model training: learning rate = 1*e* – 3, β_1 = 0.9, and β_2 = 0.999. The same random seed was used to initialize the parameters of all models. Models were trained using a single NVIDIA GeForce RTX 3060 Ti GPU. Each model took approximately one hour to train.

For a small fraction of attempted model training runs (10/85 = 11.8%), the model either failed to exceed chance accuracy (3/10 models) or converged to a state in which almost all trials had exclusively negative drift rates (7/10 models). These models were not included in summary analyses.

Convolutional neural network (CNN)

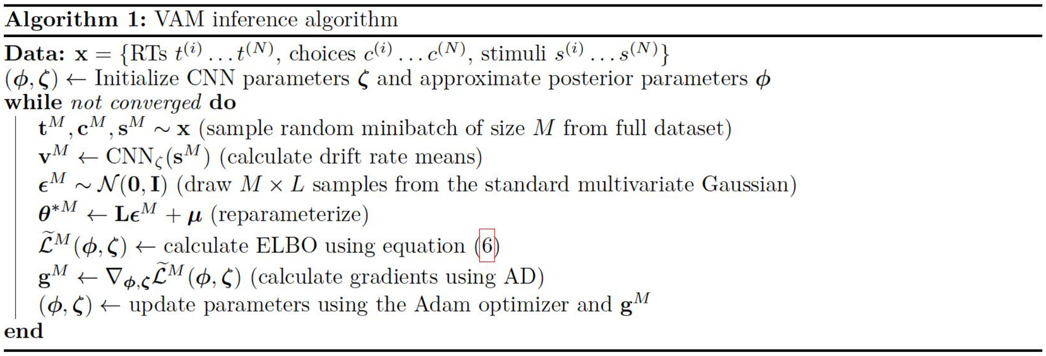
We used the same seven-layer CNN architecture for all models (six convolutional layers followed by one fully-connected hidden layer). The number of channels/units used in the seven layers was as follows: 64, 64, 128, 128, 128, 256, 1024. The output layer (after the fully-connected layer) has four channels, one for each drift rate. Each convolutional layer used a 3×3 pixel kernel (stride 1, same padding). Each convolutional layer was followed by a ReLU nonlinearity, instance normalization [89], and a max-pooling layer (2×2 pixel window, stride 2), in that order. The fully-connected hidden layer was followed by a ReLU nonlinearity and a dropout layer (dropout rate set to 0.5).

To speed up model training, the first two convolutional layers were initialized with the parameters from a larger CNN trained on an image classification task. Specifically, the pretrained model used a 16-layer VGG architecture and was trained on the ImageNet dataset [13], [79]. These first two layers were trainable (i.e., not fixed to their initial values). The weights of the other convolutional layers, fully-connected layer and output layer were initialized randomly using the LeCun normal initializer [41]; the biases were initialized with zeros.

Image preprocessing and data augmentation

The image stimuli used to train the VAM differed from the original Lost in Migration stimuli in a few ways. Information about the current score and time remaining visible at the top of the game window was removed. We also used the same blue background image for all stimuli (the background in the original Lost in Migration changes over the course of the gameplay). Finally, the stimuli were resized from 640×480 pixels (width×height) to 128×128 pixels.

We used data augmentation techniques commonly used in other image classification training paradigms to improve the generalization ability of the models. Specifically, for each batch of training data, each image was independently and randomly translated by a small amount. The size of the translation (in pixels) was drawn from a uniform distribution on the interval [0,1] (vertical)



```
**Algorithm 1:** VAM inference algorithm
**Data:** $\mathbf{x} = \{\text{RTs } t^{(i)} \ldots t^{(N)}, \text{ choices } c^{(i)} \ldots c^{(N)}, \text{ stimuli } s^{(i)} \ldots s^{(N)}\}$
$(\boldsymbol{\phi}, \boldsymbol{\zeta}) \leftarrow$ Initialize CNN parameters $\boldsymbol{\zeta}$ and approximate posterior parameters $\boldsymbol{\phi}$
**while** *not converged* **do**
    $\mathbf{t}^M, \mathbf{c}^M, \mathbf{s}^M \sim \mathbf{x}$ (sample random minibatch of size $M$ from full dataset)
    $\mathbf{v}^M \leftarrow \text{CNN}_{\zeta}(\mathbf{s}^M)$ (calculate drift rate means)
    $\boldsymbol{\epsilon}^M \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ (draw $M \times L$ samples from the standard multivariate Gaussian)
    $\boldsymbol{\theta}^{*M} \leftarrow \mathbf{L}\boldsymbol{\epsilon}^M + \boldsymbol{\mu}$ (reparameterize)
    $\widetilde{\mathcal{L}}^M(\boldsymbol{\phi}, \boldsymbol{\zeta}) \leftarrow$ calculate ELBO using equation (6)
    $\mathbf{g}^M \leftarrow \nabla_{\boldsymbol{\phi}, \boldsymbol{\zeta}} \widetilde{\mathcal{L}}^M(\boldsymbol{\phi}, \boldsymbol{\zeta})$ (calculate gradients using AD)
    $(\boldsymbol{\phi}, \boldsymbol{\zeta}) \leftarrow$ update parameters using the Adam optimizer and $\mathbf{g}^M$
**end**
```

Algorithm 1

VAM inference algorithm

and [0,2] (horizontal), rounded to the nearest pixel (horizontal and vertical translations were sampled independently). We also applied a variation of a random elastic image deformation [77] as implemented in the Augmax Python package [28]. Specifically, we used the Warp transformation in Augmax and set the strength parameter to 3 and the coarseness parameter to 32. The transformation was applied to each image independently with a probability of 0.75.

Task-optimized models

The task-optimized CNN models were trained by minimizing the cross-entropy loss function, where the correct label was defined as the true direction of the target bird in each stimulus (i.e., these models were not trained to match the decisions of the human participants, but rather to output the correct target direction). Other than the loss function, all aspects of the training process were the same as those used for the VAM (CNN architecture, optimizer settings, initialization, etc.). For each VAM we trained, we trained one task-optimized model using the same training data (n = 75 task-optimized models).

Analysis methods: overview

We analyzed all of the VAMs after 12,800 parameter update steps (corresponding to the 200th training epoch), by which point training had converged. We analyzed all of the task-optimized models after 1,600 parameter update steps (the 25th training epoch) since these models converged much more quickly. All analyses were conducted using a holdout set of n = 5000 LIM stimuli/RTs/choices (i.e., data that was not used to train the models). To generate RTs and choices from the trained models, the LIM stimuli from the holdout set were provided as inputs to the CNN, which output mean drift rates for each stimulus. We denote the drift rate mean for the k^{th} accumulator on trial i as $v_k^{(i)}$. We used the LBA model to sample RTs and choices using these mean drift rates, where the drift rate for the k^{th} accumulator on trial i, $d_k^{(i)}$, is sampled from a normal distribution with mean $v_k^{(i)}$ and SD = 1. For a very small fraction of trials (mean ± s.e.m. percent excluded trials: 0.080 ± 0.011%, n = 75 models), all of the sampled drift rates were negative, and thus the RT and choice were undefined. These trials were excluded from all analyses.

All error bars shown in the figures correspond to bootstrap 95% confidence intervals calculated using 1000 bootstrap samples.

Behavior analyses

The RT congruency effect was defined as the mean RT on incongruent trials minus the mean RT on congruent trials, where only correct trials were included in the calculation. The accuracy congruency effect was defined as the accuracy on congruent trials minus the accuracy on incongruent trials.

To calculate the RT delta plots for a given model/participant, we first calculated the RT deciles (0.1, 0.2, …, 0.9 quantiles) separately for congruent and incongruent trials, using only correct trials. Within each RT decile, we then calculated the RT congruency effect and mean RT (average of congruent and incongruent mean RTs), forming a delta plot for that participant/model. These mean RTs and congruency effects were then averaged across participants.

To calculate the conditional accuracy functions for a given model/participant, we calculated the accuracy and mean RT of trials within each RT quintile (0.2, 0.4, 0.6, and 0.8 quantiles) separately for congruent and incongruent trials. As above, these measures were then averaged across participants.

The models/participants included in the analysis of how RT varied with stimulus layout and horizontal/vertical position were those that exhibited significant modulation of RT by one or more of these stimulus features. Significant modulation was determined by running an ANOVA on the RTs in each stimulus feature bin (e.g., for each layout or each horizontal position bin), using a threshold *p*-value of 0.05. For horizontal stimulus position, we used bins of width = 50 pixels in the original 640×480 pixel window space, except for the leftmost and rightmost bins which had width = 60 pixels. For vertical stimulus position, we used bins of width = 25 pixels.

The number of participants exhibiting significant modulation of RT was 60 for stimulus layout, 72 for horizontal position, and 69 for vertical position (out of 75 total). To determine whether a given participant exhibited significant modulation of accuracy by a given stimulus feature, we used a chi-squared test for equality of proportions across stimulus feature bins (threshold *p*-value = 0.05). No participants exhibited significant modulation of accuracy for any of the stimulus features.

LBA parameters

For analyses of the fitted LBA parameters t_0, b, and A, we used the maximum a posteriori (MAP) estimates of these parameters, corresponding to the mean parameter vector of the learned Gaussian posterior density $q(\boldsymbol{\theta}; \boldsymbol{\varphi})$. For analyses of the mean target and flanker drift rates, we provided the LIM stimuli from the holdout set as inputs to the fitted CNN, which generates the mean drift rates as its outputs. The non-target/non-flanker (other) drift rates were calculated by averaging the drift rates from the two non-target/non-flanker accumulators on incongruent trials or the three non-target accumulators on congruent trials.

CNN unit activity

The activation matrices used in the analyses of CNN representations were derived from the responses of units in each layer elicited by the holdout image set. The activations were processed just after the ReLU nonlinearity. For the convolutional layers, we defined the activation of a given unit/channel as the maximum value of that channel calculated across the spatial dimensions [29]. This yields a $N \times K_l$ activation matrix, where N is the number of images in the holdout set (5000) and K_l is the number of active channels in layer l. For the fully-connected layer, the responses of all units were used to define the $N \times K_l$ activation matrix directly.

Only incongruent trials were used for all of the analyses involving the CNN activations described below, for the following reasons. For the selectivity and tolerance analyses in which we decoded target or flanker direction from the activity in each layer, described in more detail below, note that the target and flanker directions are by definition the same on congruent trials. As such, a classifier trained to decode target direction from congruent trials could achieve perfect accuracy using information in the unit activity attributable to the flankers, rather than targets. Using only incongruent trials ensured that such cross-contamination could not occur. We used only incongruent trials for all of the other analyses of the activations simply for convenience.

Stimulus feature decoding

To assess how well particular stimulus features could be decoded from the activity in each layer, we trained a linear SVM to classify the values of that stimulus feature using the $N \times K_l$ activation matrix in each layer. The activation matrix was standardized before training the classifiers, such that each column had zero mean and unit variance. Horizontal and vertical stimulus positions were discretized to enable classification using the same bins as were used in the behavioral analyses.

The classification task was done in a standard "one-vs-rest" setting: for each value of a given stimulus feature, one sub-classifier was trained on a binary classification task with the chosen value as one class and all other values as the other class, yielding one classifier for each value of the stimulus feature (e.g., four for target direction, seven for stimulus layout). To determine the decision of the combined classifier for a given image, we generated predictions from each sub-classifier and assigned the decision to the sub-classifier with the largest (most confident) prediction. We assessed the overall decoding accuracy of each SVM on a separate test image set. Note that both the training set and test set for the SVMs were subsets of the holdout image set (i.e., the images that were not used to train the VAMs). The SVMs were trained using the LinearSVC model in the scikit-learn Python package [60] with the squared hinge loss function and L2 regularization with the penalty parameter C set to 1.0 (the default settings).

Tolerance

To assess the tolerance of model representations to variation in a given stimulus feature (flanker direction, stimulus layout, horizontal position, and vertical position), we trained a linear SVM to classify target direction from the CNN activations using stimuli with the chosen stimulus feature fixed to one value (the training context), and assessed the generalization performance of the SVM on stimuli that contained all other values of that stimulus feature (the generalization context). For example, to assess tolerance to stimulus layout, we trained one SVM to classify target direction using stimuli with the vertical line layout, and assessed the generalization performance of that classifier on stimuli with the six other layouts. We trained one such SVM for each of the seven stimulus layouts, and averaged the generalization performance across these seven SVMs to derive the overall generalization performance measure for a given model and network layer. Other details of the classifiers were the same as those used for the decoding analyses described above.

Target/flanker subspace alignment

To calculate the target/flanker subspace alignment metric, we first defined target and flanker subspaces using the SVM classifiers that we trained for the target/flanker decoding analyses. Specifically, for each of the four classifiers, which were trained to classify stimuli as a given target or flanker direction vs. all other target or flanker directions using the CNN activations from a given layer, we extracted the vector orthogonal to the decision hyperplane. In agreement with prior work [5], [46], we refer to these vectors as decoding vectors for a given target/flanker direction. Let $\mathbf{x}_{k,\text{targ}}^T$ and $\mathbf{x}_{k,\text{flnk}}^T$ denote the target decoding row vector and flanker decoding row vector for the kth direction, respectively. We define the matrices formed with these four decoding vectors filling the rows as the target subspace matrix $\mathbf{X}_{\text{targ}}$ and the flanker subspace matrix $\mathbf{X}_{\text{flnk}}$. These matrices have dimensions 4 × K_l, where K_l is the number of active units in layer l. Each matrix therefore spans a subspace of the full K_l-dimensional space. Our goal is to determine whether the target and flanker subspaces are orthogonal.

To do so, we employ principal angles between subspaces [36], [99], which generalizes the more intuitive notion of angles between lines or planes to arbitrary dimensions. To calculate the principal angles, we require orthonormal bases for the target and flanker subspaces, which we determine using a reduced singular value decomposition (SVD):

$$\mathbf{X}_{\text{targ}} = \mathbf{U}_{\text{targ}}\mathbf{\Sigma}_{\text{targ}}\mathbf{V}_{\text{targ}}^T, \qquad \mathbf{X}_{\text{flnk}} = \mathbf{U}_{\text{flnk}}\mathbf{\Sigma}_{\text{flnk}}\mathbf{V}_{\text{flnk}}^T.$$

The rows of the 4 × K_l matrices $\mathbf{V}_{\text{targ}}^T$ and $\mathbf{V}_{\text{flnk}}^T$ form an orthonormal basis for the target and flanker subspaces, respectively. The cosines of the principal angles are given by the singular values of $\mathbf{V}_{\text{targ}}\mathbf{V}_{\text{flnk}}^T$. The average of these singular values is our subspace alignment metric, which ranges from zero (completely orthogonal subspaces) to one (completely aligned/parallel subspaces).

Participation ratio

We measured the dimensionality of target representations for a given layer with the participation ratio (PR_l) [23], defined as:

$$\mathrm{PR}_l = \frac{\left(\sum_{i=1}^{K_l} \lambda_i\right)^2}{\sum_{i=1}^{K_l} \lambda_i^2},$$

where K_l is the number of active units in layer $\lambda_1 \geq \ldots \geq \lambda_i \geq \ldots \geq \lambda_{K_l}$ are the eigenvalues of the target-centered activation covariance matrix for layer l. The target-centered activations were obtained by subtracting the centroid of the activation matrix for each target direction from the corresponding trials in the activation matrix [64].

The participation ratio is a continuous measure of dimensionality ranging from 1 to K_l. The minimum ($\mathrm{PR}_l = 1$) is obtained when all of the variance in activity is concentrated in a single dimension, such that $\lambda_i = 0$ for $i \geq 2$. The maximum ($\mathrm{PR}_i = K_l$) is obtained when the variance is evenly spread across the K_l dimensions, such that all K_l eigenvalues are equal.

Mutual information

The activity of each unit in response to the holdout image set was discretized into 10 equally-sized bins, yielding a unit-specific activation distribution pX (x). Stimulus features (horizontal/vertical position, layout, target/flanker direction) were discretized if the values were continuous, or used as is if not, yielding a stimulus feature distribution pY (y). Discretization of the continuous variables was done as described in the decoding methods section. The joint probability mass function of the unit activity and stimulus feature is denoted by $p(X,Y)^{(x,\, y)}$. For a given unit and stimulus feature, the mutual information is given by:

$$I(X;Y) = \sum_{x,y} p_{(X,Y)}(x,y) \log \frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)}.$$

The mutual information for a given stimulus feature was normalized by the entropy of the feature distribution to facilitate comparisons between the features [53]. The entropy is given by:

$$H(Y) = -\sum_{y} p_Y(y) \log p_Y(y).$$

Single-unit modulation by target direction

To identify units that were modulated by target direction, we ran a one-way ANOVA on the z-scored activity of each unit, where each group in the ANOVA was determined by the activity of that unit in response to stimuli for a given target direction (the activity of each unit was z-scored across the stimuli). Units with an ANOVA p-value < 0.001 were defined as significantly modulated by target direction. These units were further split into three subtypes based on their degree of selectivity and sign of modulation: selective (+), selective (-), and complex units.

We first selected units that had significantly higher or lower activation for one direction relative to the other three directions, as assessed by Tukey's HSD test ($p < 0.05$). Within this population, some units had both significantly higher activation for one direction relative to the other three *and* significantly lower activation for one direction relative to the other three. Units that did vs. did not have this property were handled separately. In the former (simpler) case, the units that only had significantly higher activation for one direction relative to the other three were defined as selective (+) units; the units that only had significantly lower activation for one direction relative to the other three were defined as selective (-) units.

In the latter (more complicated) case, we compared the magnitude of the activation for the positive and negative modulation directions with a rank-sum test. Units for which the magnitude of activity for the positive modulation direction was significantly greater ($p < 0.05$) than the magnitude of activity for the negative modulation direction were defined as selective (+) units. Analogous criteria were used to define selective (-) units (with the signs reversed). Units within this pool with rank-sum p-value > 0.05 were defined as complex units. Finally, units that were significantly modulated by target direction but that did not meet any of the criteria described above were also defined as complex units.

Supporting Information

Data and code availability

All of the code (*https://github.com/pauljaffe/vam*) and data (*https://doi.org/10.5281/zenodo.10775514*) used to train the VAMs and reproduce our results are publicly-available without restrictions.

Acknowledgements

We thank M. Steyvers, M. Robinson, D. Yamins and members of the NeuroAILab, the Lumos Labs research team, and members of the Poldrack Lab for helpful conversations and comments on this work. We also thank A. Kaluszka for providing graphics files of the Lost and Migration stimuli. The authors received no specific funding for this work.

Additional information

Author contributions

P.I.J. designed research, performed research, and wrote the paper. P.I.J. and G.X.S.R. analyzed data. P.I.J., G.X.S.R., R.J.S., P.G.B., and R.A.P. edited the paper. R.J.S. and R.A.P. provided resources. R.J.S., P.G.B., and R.A.P. supervised research and provided input at all stages.

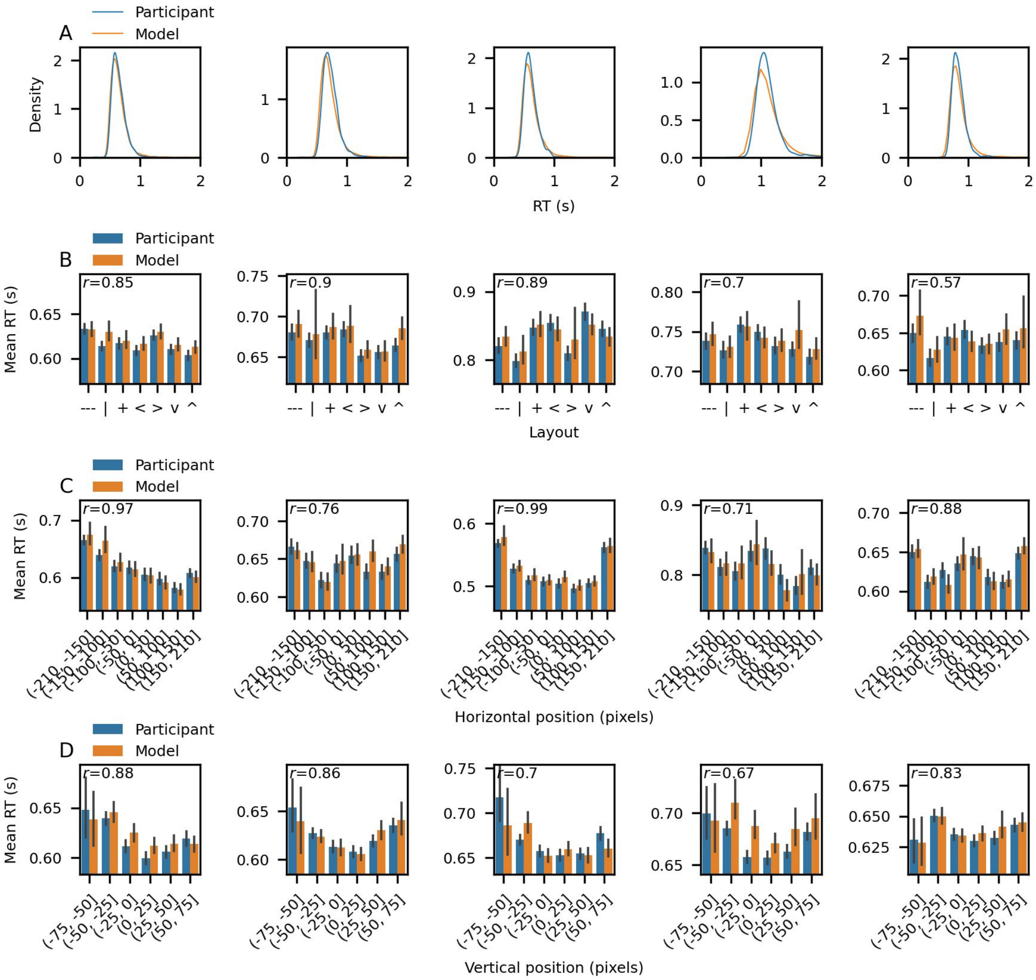


Fig. S1

Example model/participant RT distributions and dependence of RTs on stimulus features.

A) Example model/participant RT distributions (all trials). **B)** Examples of model/participant mean RT vs. stimulus layout. **C)** Examples of model/participant mean RT vs. horizontal stimulus position (negative values: left of center). **D)** Examples of model/participant mean RT vs. vertical stimulus position (negative values: above center). For all panels, error bars correspond to bootstrap 95% confidence intervals.

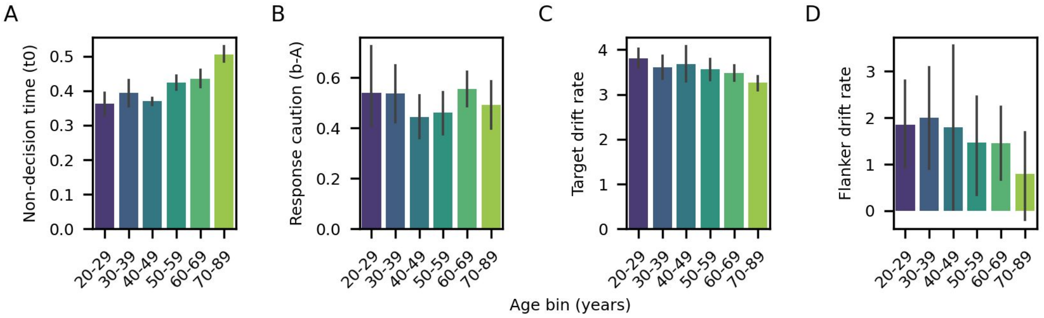


Fig. S2

Age dependence of LBA parameters.

For all panels, we tested age-dependence with a one-way ANOVA and report Bonferroni-adjusted *p*-values, corrected for 4 comparisons ($n = 75$ models). We also report adjusted *p*-values from a post-hoc comparison of the 20-29 vs. 70-89 age groups conducted with Tukey's HSD. Error bars correspond to bootstrap 95% confidence intervals. **A)** Non-decision time parameter t_0 ($F\,(5, 69) = 13.3$, $p < 1e\text{-}7$). Tukey's HSD for 20-29 vs. 70-89 age groups: $p < 1e\text{-}8$. **B)** Response caution ($b - A$; $F\,(5, 69) = 0.49$, $p = 1.0$). **C)** Mean target drift rate ($F\,(5, 69) = 3.4$, $p = 0.026$). Tukey's HSD for 20-29 vs. 70-89 age groups: $p = 0.002$. **D)** Mean flanker drift rate ($F\,(5, 69) = 0.72$, $p = 1.0$).

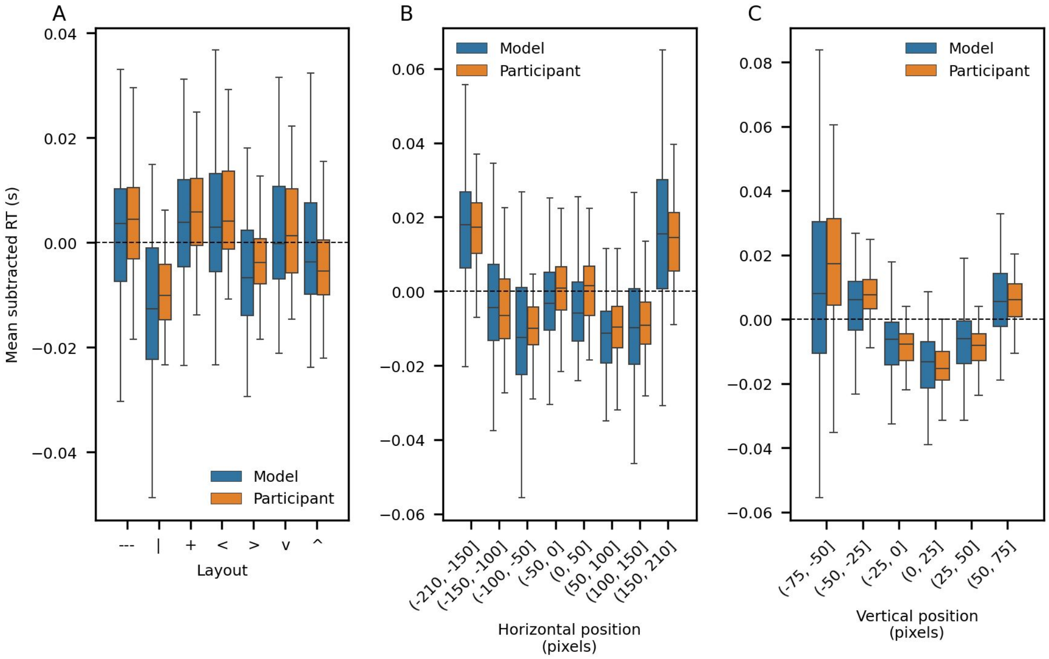


Fig. S3

Dependence of RTs on stimulus layout and position.

For each participant/model, we calculated the mean RT in each stimulus feature bin, then subtracted the average of these mean RTs from each bin. The panels show the average of these centered RTs across all participants with significant modulation of RT for that particular stimulus feature. For all panels, we conducted a one-way ANOVA for both models/participants and report Bonferroni-adjusted p-values, corrected for 3 comparisons (n = 75 models). We also report results from post-hoc comparisons between select feature bins conducted with Tukey's HSD. Error bars correspond to bootstrap 95% confidence intervals. **A)** RT vs. stimulus layout (models: F (6, 53) = 7.43, p < 1e-6, RTs for the vertical line layout were significantly faster (Tukey's HSD adjusted p-value < 0.05) than RTs from all other layouts except '>'; participants: F (6, 53) = 23.1, p < 1e-22; RTs for the vertical line layout were significantly faster than RTs from all other layouts). **B)** RT vs. horizontal stimulus position (negative values: left of center; models: F (7, 64) = 16.8, p < 1e-18, RTs for the leftmost and rightmost position bins were significantly slower than RTs from all intermediate position bins; participants: F (7, 64) = 72.6, p < 1e-73; RTs for the leftmost and rightmost position bins were significantly slower than RTs from all intermediate position bins). **C)** RT vs. vertical stimulus position (negative values: above center; models: F (5, 66) = 17.2, p < 1e-14, RTs for the topmost and bottommost position bins were significantly slower than RTs from the two centermost position bins; participants: F (5, 66) = 113.3, p < 1e-74; RTs for the topmost and bottommost position bins were significantly slower than RTs from the two centermost position bins).

Fig. S4

RT delta plots and conditional accuracy functions.

A) RT delta plots for participants and VAMs (*n* = 75 models/participants). **B)** Conditional accuracy functions for participants and VAMs. For all panels, error bars correspond to bootstrap 95% confidence intervals.

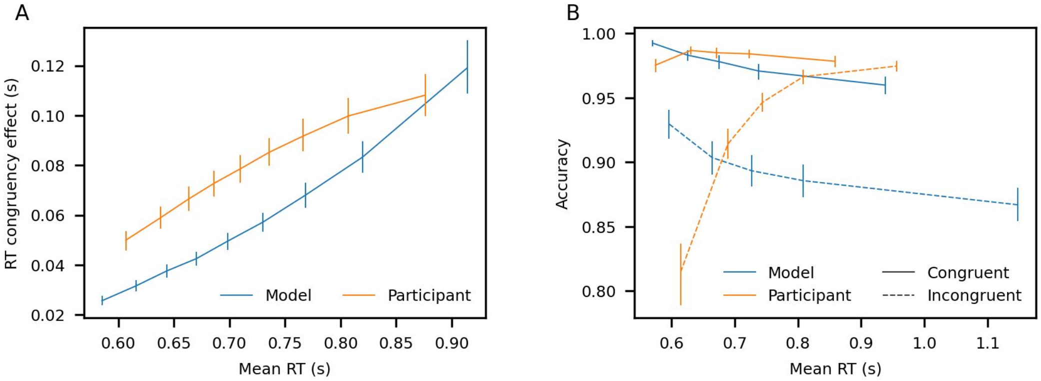


Fig. S5

Activity of all selective (+) units for one example model.

Each row shows the activity of one unit for 100 randomly selected stimuli, sorted by target direction. The activity of each unit was centered and normalized by the activity of the stimulus with the largest magnitude activation. The small number of selective (+) units in layer Conv1 are not shown.

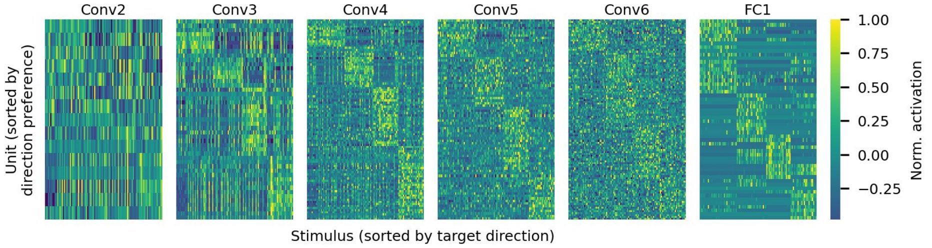


Fig. S6

Absence of correlation between flanker suppression metrics and congruency effects.

All panels show the Pearson’s correlation coefficient between the specified suppression and behavior metrics, calculated across models. **A)** Flanker direction decoding accuracy vs. accuracy congruency effect. **B)** Mutual information for flanker direction conveyed by single units vs. accuracy congruency effect. **C)** Flanker direction decoding accuracy vs. RT congruency effect. **D)** Mutual information for flanker direction conveyed by single units vs. RT congruency effect. For all panels, *n* = 75 models, error bars correspond to bootstrap 95% confidence intervals. Asterisks indicate a significant Pearson’s *r* (adjusted *p*-value < 0.05, permutation test with *n* = 1000 shuffles, Bonferroni correction for 7 comparisons).

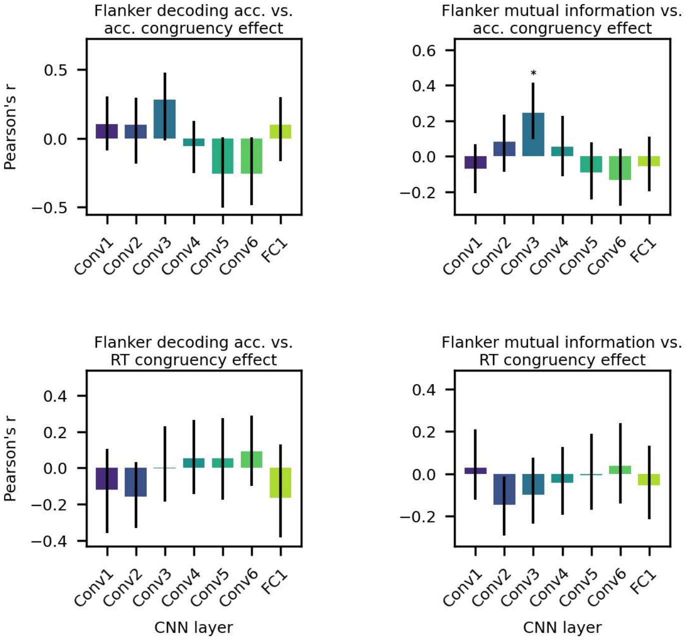
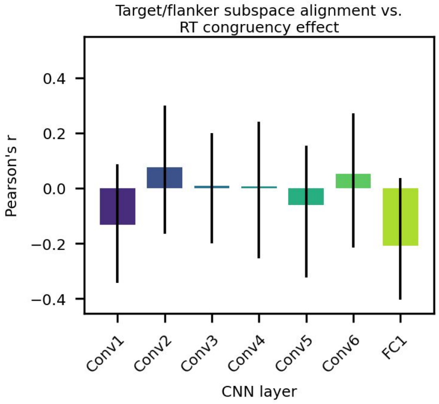


Fig. S7

Absence of correlation between target/flanker subspace alignment and RT congruency effect.

Pearson’s correlation coefficient between target/flanker subspace alignment and RT congruency effect across models (*n* = 75 models, error bars correspond to bootstrap 95% confidence intervals). The correlation was not significant for any layer (adjusted *p*-value > 0.05, permutation test with *n* = 1000 shuffles, Bonferroni correction for 7 comparisons).



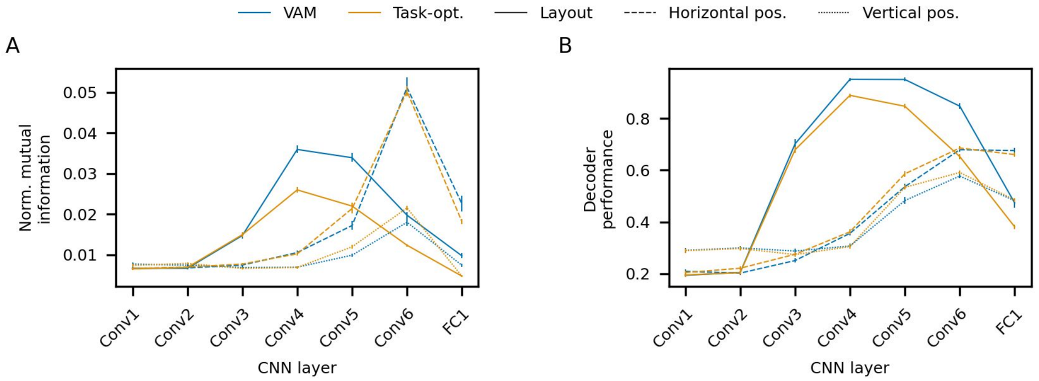


Fig. S8

Additional analysis of VAMs and task-optimized models.

A) Normalized mutual information for stimulus layout and horizontal/vertical stimulus position conveyed by single units, averaged across units. Mutual information was normalized by the entropy of the corresponding stimulus feature distribution. **B)** Decoding accuracy of stimulus layout and horizontal/vertical stimulus position. All panels show the average across *n* = 75 task-optimized models and *n* = 75 VAMs; error bars correspond to bootstrap 95% confidence intervals. The VAM data shown in panels A and B is the same as that shown in **Figs. 4C and 4B**, respectively.

References

[1] Annis J., Gauthier I., Palmeri T. J. (2021) **Combining Convolutional Neural Networks and Cognitive Models to Predict Novel Object Recognition in Humans** *J. Exp. Psychol. Learn. Mem. Cogn* **47**:785–807

[2] Ansuini A., Laio A., Macke J. H., Zoccolan D. (2019) **Intrinsic dimension of data representations in deep neural networks** *Adv. Neural Inf. Process. Syst* **32**

[3] Baker N., Lu H., Erlikhman G., Kellman P. J. (2018) **Deep convolutional networks do not classify based on global object shape** *PLoS Comput. Biol* **14**

[4] Ben-David B. M., Eidels A., Donkin C. (2014) **Effects of Aging and Distractors on Detection of Redundant Visual Targets and Capacity: Do Older Adults Integrate Visual Targets Differently than Younger Adults?** *PLoS One* **9**

[5] Bernardi S., Benna M. K., Rigotti M., Munuera J., Fusi S., Salzman C. D. (2020) **The geometry of abstraction in the hippocampus and prefrontal cortex** *Cell* **183**:954–967

[6] Bowers J. S., Malhotra G., Dujmović M., et al. (2022) **Deep Problems with Neural Network Models of Human Vision** *Behav. Brain Sci* :1–74

[7] Bradbury J., Frostig R., Hawkins P., et al. (2018) **JAX: Composable transformations of Python+NumPy programs** *GitHub* http://github.com/google/jax

[8] Brincat S. L., Siegel M., Nicolai C. v., Miller E. K. (2018) **Gradual progression from sensory to task-related processing in cerebral cortex** *Proc. Natl. Acad. Sci. U. S. A* **115**

[9] Brown S. D., Heathcote A. (2008) **The simplest complete model of choice response time: Linear ballistic accumulation** *Cogn. Psychol* **57**:153–178

[10] Cohen J. D., Servan-Schreiber D., McClelland J. L. (1992) **A Parallel Distributed Processing Approach to Automaticity** *Am. J. Psychol* **105**

[11] Cover T. M. (1965) **Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition** *IEEE Trans. Electron. Comput* :326–334

[12] Dao V. H., Gunawan D., Tran M.-N., Kohn R., Hawkins G. E., Brown S. D. (2022) **Efficient Selection Between Hierarchical Cognitive Models: Cross-Validation With Variational Bayes** *Psychol. Methods*

[13] Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L. (2009) **ImageNet: A large-scale hierarchical image database** *Proc. IEEE Conf. Comput. Vis. Pattern Recognit* :248–255

[14] Dezfouli A., Griffiths K., Ramos F., Dayan P., Balleine B. W. (2019) **Models that learn how humans learn: The case of decision-making and its disorders** *PLoS Comput. Biol* **15**

[15] DiCarlo J. J., Zoccolan D., Rust N. C. (2012) **How Does the Brain Solve Visual Object Recognition?** *Neuron* **73**:415–434

[16] Dosovitskiy A., Beyer L., Kolesnikov A., et al. (2021) **An image is worth 16x16 words: Transformers for image recognition at scale** *International Conference on Learning Representations*

[17] Eckstein M. P., Koehler K., Welbourne L. E., Akbas E. (2017) **Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes** *Curr. Biol* **27**:2827–2832

[18] Eriksen B. A., Eriksen C. W. (1974) **Effects of noise letters upon the identification of a target letter in a nonsearch task** *Percept. Psychophys* **16**:143–149

[19] Evans N. J., Wagenmakers E.-J. (2020) **Evidence accumulation models: Current limitations and future directions** *Quant. Meth. Psychol* **16**:73–90

[20] Fel T., Rodriguez I. F. Rodriguez, Linsley D., Serre T. (2022) **Harmonizing the object recognition strategies of deep neural networks with humans** *Adv. Neural Inf. Process. Syst.,* **35**:9432–9446

[21] Flesch T., Juechems K., Dumbalska T., Saxe A., Summerfield C. (2022) **Orthogonal representations for robust context-dependent task performance in brains and neural networks** *Neuron* **110**:1258–1270

[22] Forstmann B. U., Tittgemeyer M., Wagenmakers E.-J., Derrfuss J., Imperati D., Brown S. (2011) **The Speed-Accuracy Tradeoff in the Elderly Brain: A Structural Model-Based Approach** *J. Neurosci* **31**:17242–17249

[23] Gao P., Trautmann E., Yu B., et al. (2017) **A theory of multineuronal dimensionality, dynamics and measurement** *bioRxiv* https://doi.org/10.1101/214262

[24] Goetschalckx L., Govindarajan L. N., Karkada Ashok A., Ahuja A., Sheinberg D., Serre T. (2023) **Computing a human-like reaction time metric from stable recurrent vision models** *Adv. Neural Inf. Process. Syst* **36**:14338–14365

[25] Gottsdanker R. (1982) **Age and Simple Reaction Time** *J. Gerontol* **37**:342–348

[26] Güçlü U., van Gerven M. A. J. (2015) **Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream** *J. Neurosci* **35**:10005–10014

[27] Gunawan D., Hawkins G., Tran M.-N., Kohn R., Brown S. (2020) **New estimation approaches for the hierarchical linear ballistic accumulator model** *J. Math. Psychol* **96**

[28] Heidler K. (2022) **Augmax** *GitHub* https://github.com/khdlr/augmax

[29] Hohman F., Park H., Robinson C., Polo Chau D. H. (2020) **Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations** *IEEE Trans. Vis. Comput. Graph* **26**:1096–1106

[30] Holmes W. R., O'Daniels P., Trueblood J. S. (2020) **A Joint Deep Neural Network and Evidence Accumulation Modeling Approach to Human Decision-Making with Naturalistic Images** *Comput. Brain Behav* **3**:1–12

[31] Hung C. P., Kreiman G., Poggio T., DiCarlo J. J. (2005) **Fast Readout of Object Identity from Macaque Inferior Temporal Cortex** *Science* **310**:863–866

[32] Jacobs R. A., Bates C. J. (2019) **Comparing the Visual Representations and Performance of Humans and Deep Neural Networks** *Curr. Dir. Psychol* **28**:34–39

[33] Jaffe P. I., Poldrack R. A., Schafer R. J., Bissett P. G. (2023) **Modelling human behaviour in cognitive tasks with latent dynamical systems** *Nat. Hum. Behav* **7**:986–1000

[34] Jha A., Peterson J. C., Griffiths T. L. (2023) **Extracting Low-Dimensional Psychological Representations from Convolutional Neural Networks** *Cogn. Sci* **47**

[35] Jong R. D., Liang C.-C., Lauber E. (1994) **Conditional and Unconditional Automaticity: A Dual-Process Model of Effects of Spatial Stimulus-Response Correspondence** *J. Exp. Psychol. Hum. Percept. Perform* **20**:731–750

[36] Jordan C. (1875) **Essai sur la géométrie à n dimensions** *fr, Bulletin de la Société Mathématique de France* **3**:103–174

[37] Kaufman M. T., Churchland M. M., Ryu S. I., Shenoy K. V. (2014) **Cortical activity in the null space: permitting preparation without movement** *Nat. Neurosci* **17**:440–448

[38] Kingma D. P., Welling M. (2013) **Auto-Encoding Variational Bayes** *arXiv*

[39] Kingma D. P., Ba J. (2017) **Adam: A method for stochastic optimization** *arXiv*

[40] Kingma D. P., Salimans T., Welling M. (2015) **Variational dropout and the local reparameterization trick** *arXiv*

[41] Klambauer G., Unterthiner T., Mayr A., Hochreiter S. (2017) **Self-normalizing neural networks** *arXiv*

[42] Koren V., Andrei A. R., Hu M., Dragoi V., Obermayer K. (2020) **Pairwise synchrony and correlations depend on the structure of the population code in visual cortex** *Cell Rep* **33**

[43] Kriegeskorte N. (2015) **Deep neural networks: A new framework for modeling biological vision and brain information processing** *Annu. Rev. Vis. Sci* **1**:417–446

[44] Kucukelbir A., Tran D., Ranganath R., Gelman A., Blei D. M. (2017) **Automatic differentiation variational inference** *J. Mach. Learn. Res* **18**:1–45

[45] Kumbhar O., Sizikova E., Majaj N., Pelli D. G. (2020) **Anytime Prediction as a Model of Human Reaction Time** *arXiv*

[46] Libby A., Buschman T. J. (2021) **Rotational dynamics reduce interference between sensory and memory representations** *Nat. Neurosci.,* **24**:715–726

[47] Lindsay G. W. (2021) **Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future** *J. Cogn. Neurosci* **33**:2017–2031

[48] Linsley D., Eberhardt S., Sharma T., Gupta P., Serre T. (2017) **What are the Visual Features Underlying Human Versus Machine Vision?** *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)* :2706–2714

[49] Lo C.-F., Ip H.-Y. (2021) **Modified leaky competing accumulator model of decision making with multiple alternatives: the Lie-algebraic approach** *Sci. Rep* **11**

[50] Malhotra G., Dujmović M., Bowers J. S. (2022) **Feature blindness: A challenge for understanding and modelling visual object recognition** *PLoS Comput. Biol* **18**

[51] Mante V., Sussillo D., Shenoy K. V., Newsome W. T. (2013) **Context-dependent computation by recurrent dynamics in prefrontal cortex** *Nature* **503**:78–84

[52] Meister M. L. R., Hennig J. A., Huk A. C. (2013) **Signal Multiplexing and Single-Neuron Computations in Lateral Intraparietal Area During Decision-Making** *J. Neurosci* **33**:2254–2267

[53] Muratore P., Tafazoli S., Piasini E., Laio A., Zoccolan D. (2022) **Prune and distill: Similar reformatting of image information along rat visual cortex and deep neural networks** *Adv. Neural Inf. Process. Syst*

[54] Navarro D. J., Fuss I. G. (2009) **Fast and accurate calculations for first-passage times in wiener diffusion models** *J. Math. Psychol* **53**:222–230

[55] Nayebi A., Bear D., Kubilius J., et al. (2018) **Task-Driven Convolutional Recurrent Models of the Visual System** *arXiv*

[56] Nettelbeck T., Rabbitt P. M. (1992) **Aging, cognitive performance, and mental speed** *Intelligence* **16**:189–205

[57] Pagan M., Urban L. S., Wohl M. P., Rust N. C. (2013) **Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information** *Nat. Neurosci* **16**:1132–1139

[58] Panichello M. F., Buschman T. J. (2021) **Shared mechanisms underlie the control of working memory and attention** *Nature* **592**:601–605

[59] Papyan V., Han X. Y., Donoho D. L. (2020) **Prevalence of neural collapse during the terminal phase of deep learning training** *Proc. Natl. Acad. Sci. U. S. A* **117**:24652–24663

[60] Pedregosa F., Varoquaux G., Gramfort A., et al. (2011) **Scikit-learn: Machine learning in Python** *J. Mach. Learn. Res* **12**:2825–2830

[61] Pratte M. S. (2021) **Eriksen flanker delta plot shapes depend on the stimulus** *Atten. Percept. Psychophys* **83**:685–699

[62] Rafiei F., Shekhar M., Rahnev D. (2024) **The neural network RTNet exhibits the signatures of human perceptual decision-making** *Nat. Hum. Behav* **8**:1752–1770

[63] Rajalingham R., Issa E. B., Bashivan P., Kar K., Schmidt K., DiCarlo J. J. (2018) **Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks** *J. Neurosci* **38**:7255–7269

[64] Rangamani A., Lindegaard M., Galanti T., Poggio T. A. (2023) **Feature learning in deep classifiers through Intermediate Neural Collapse** *Proc. Mach. Learn. Res* **202**:28729–28745

[65] Ratcliff R., Thapar A., McKoon G. (2001) **The effects of aging on reaction time in a signal detection task** *Psychol. Aging* **16**

[66] Ratcliff R. (1978) **A theory of memory retrieval** *Psychol. Rev* **85**:59–108

[67] Ratcliff R., McKoon G. (2008) **The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks** *Neural Comput* **20**:873–922

[68] Ratcliff R., Rouder J. N. (1998) **Modeling response times for two-choice decisions** *Psychol. Sci* **9**:347–356

[69] Rezende D. J., Mohamed S., Wierstra D. (2014) **Stochastic Backpropagation and Approximate Inference in Deep Generative Models** *arXiv*

[70] Ridderinkhof K. R. (2002) **Activation and suppression in conflict tasks: Empirical clarification through distributional analyses** *Common Mechanisms in Perception and Action: Attention and Performance XIX* Oxford University Press

[71] Ridderinkhof R. K. (2002) **Micro- and macro-adjustments of task set: Activation and suppression in conflict tasks** *Psychol. Res.,* **66**:312–323

[72] Rigotti M., Barak O., Warden M. R., et al. (2013) **The importance of mixed selectivity in complex cognitive tasks** *Nature* **497**:585–590

[73] Ritz H., Shenhav A. (2024) **Orthogonal neural encoding of targets and distractors supports multivariate cognitive control** *Nat. Hum. Behav* :1–17

[74] Rust N. C., DiCarlo J. J. (2010) **Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT** *J. Neurosci* **30**:12978–12995

[75] Sanders C. A., Nosofsky R. M. (2020) **Training Deep Networks to Construct a Psychological Feature Space for a Natural-Object Category Domain** *Comput. Brain Behav* **3**:229–251

[76] Servant M., Evans N. J. (2020) **A Diffusion Model Analysis of the Effects of Aging in the Flanker Task** *Psychol. Aging* **35**:831–849

[77] Simard P., Steinkraus D., Platt J. (2003) **Best practices for convolutional neural networks applied to visual document analysis** *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings* :958–963

[78] Simon J. (1982) **Effect of an auditory stimulus on the processing of a visual stimulus under single- and dual-tasks conditions** *Acta Psychol* **51**:61–73

[79] Simonyan K., Zisserman A. (2015) **Very deep convolutional networks for large-scale image recognition** *arXiv*

[80] Spoerer C. J., Kietzmann T. C., Mehrer J., Charest I., Kriegeskorte N. (2020) **Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision** *PLoS Comput. Biol* **16**

[81] Steyvers M., Hawkins G. E., Karayanidis F., Brown S. D. (2019) **A large-scale analysis of task switching practice effects across the lifespan** *Proc. Natl. Acad. Sci. U. S. A* **116**:17735–17740

[82] Stoffels E. J., Molen M. W. v. d. (1988) **Effects of visual and auditory noise on visual choice reaction time in a continuous-flow paradigm** *Percept. Psychophys* **44**:7–14

[83] Stroop J. R. (1935) **Studies of interference in serial verbal reactions** *J. Exp. Psychol* **18**:643–662

[84] Sussillo D., Churchland M. M., Kaufman M. T., Shenoy K. V. (2015) **A neural network that finds a naturalistic solution for the production of muscle activity** *Nat. Neurosci* **18**:1025–1033

[85] Tafazoli S., Safaai H., De Franceschi G., et al. (2017) **Emergence of transformation-tolerant representations of visual objects in rat lateral extrastriate cortex** *eLife* **6**

[86] Taylor J. E. T., Shekhar S., Taylor G. W. (2021) **Neural response time analysis: Explainable artificial intelligence using only a stopwatch** *Appl. AI Lett* **2**

[87] Trueblood J. S., Eichbaum Q., Seegmiller A. C., Stratton C., O'Daniels P., Holmes W. R. (2021) **Disentangling prevalence induced biases in medical image decision-making** *Cognition* **212**

[88] Ulrich R., Schröter H., Leuthold H., Birngruber T. (2015) **Automatic and controlled stimulus processing in conflict tasks: Superimposed diffusion processes and delta functions** *Cogn. Psychol* **78**:148–174

[89] Ulyanov D., Vedaldi A., Lempitsky V. (2017) **Instance normalization: The missing ingredient for fast stylization** *arXiv*

[90] Usher M., McClelland J. L. (2001) **The Time Course of Perceptual Choice: The Leaky, Competing Accumulator Model** *Psychol. Rev* **108**:550–592

[91] Wang J., Narain D., Hosseini E. A., Jazayeri M. (2018) **Flexible timing by temporal scaling of cortical responses** *Nat. Neurosci* **21**:102–110

[92] White C. N., Ratcliff R., Starns J. J. (2011) **Diffusion models of the flanker task: Discrete versus gradual attentional selection** *Cogn. Psychol* **63**:210–238

[93] Wildenberg W. P. v. d., Wylie S. A., Forstmann B. U., Burle B., Hasbroucq T., Ridderinkhof K. R. (2010) **To Head or to Heed? Beyond the Surface of Selective Action Inhibition: A Review** *Front. Hum. Neurosci* **4**

[94] Xie Y., Hu P., Li J., et al. (2022) **Geometry of sequence working memory in macaque prefrontal cortex** *Science* **375**:632–639

[95] Yamins D. L. K., DiCarlo J. J. (2016) **Using goal-driven deep learning models to understand sensory cortex** *Nat. Neurosci.,* **19**:356–365

[96] Yamins D. L. K., Hong H., Cadieu C. F., Solomon E. A., Seibert D., DiCarlo J. J. (2014) **Performance-optimized hierarchical models predict neural responses in higher visual cortex** *Proc. Natl. Acad. Sci. U. S. A* **111**:8619–8624

[97] Yang G. R., Joglekar M. R., Song H. F., Newsome W. T., Wang X.-J. (2019) **Task representations in neural networks trained to perform many cognitive tasks** *Nat. Neurosci* **22**:297–306

[98] Zeiler M. D., Fergus R. (2013) **Visualizing and understanding convolutional networks** *arXiv*

[99] Zhu P., Knyazev A. (2013) **Angles between subspaces and their tangents** *J. Numer. Math* **21**

Author information

Paul I Jaffe
Department of Psychology, Stanford University, Stanford, United States

For correspondence: pauljaffe7@gmail.com

Gustavo X Santiago-Reyes
Department of Bioengineering, Stanford University, Stanford, United States

Robert J Schafer
Lumos Labs, San Francisco, United States

Patrick G Bissett
Department of Psychology, Stanford University, Stanford, United States

Russell A Poldrack
Department of Psychology, Stanford University, Stanford, United States

Editors

Reviewing Editor
Marius Peelen
Radboud University Nijmegen, Nijmegen, Netherlands

Senior Editor
Michael Frank
Brown University, Providence, United States of America

Reviewer #1 (Public review):

Summary:

This paper introduces a new approach for modeling human behavioral responses using image-computable models. They create a model (VAM) that is a combination of a standard CNN coupled with a standard evidence accumulation model (EAM). The combined model is then trained directly on image-level data using human behavioral responses. This approach is original and can have wide applicability. However, many of the specific findings reported are less compelling.

Strengths:

(1) The manuscript presents an original approach of fitting an image-computable model to human behavioral data. This type of approach is sorely needed in the field.
(2) The analyses are very technically sophisticated.
(3) The behavioral data are large both in terms of sample size (N=75) and in terms of trials per subject.

Weaknesses:

(1) The main advance here thus appears to be methodological rather than conceptual. It's really cool that VAMs are image computable and are also fit to human data. But what we learn about the mind or brain is perhaps more modest.
(2) In the approach here, a given stimulus is always processed in the same way through the core CNN to produce activations v_k. These v_k's are then corrupted by Gaussian noise to

produce drift rates d_k, which can differ from trial to trial even for the same stimulus. In other words, the assumption built into VAM appears to be that the drift rate variability stems entirely from post-sensory (decisional) noise. In contrast, the typical interpretation of EAMs is that the variability in drift rates is sensory. In response to this concern, the authors responded that one can imagine an additional (unmodeled) sensory process that adds variability to the drift rates. However, this process remains unmodeled. The authors motivate their paper by saying "EAMs do not explain how the visual system extracts these representations in the first place" (second sentence of the Abstract). VAM is definitely a step in this direction but there's still a gap between the current VAM implementation and sensory systems.

https://doi.org/10.7554/eLife.98351.2.sa2

Reviewer #2 (Public review):

In An image-computable model of speeded decision-making, the authors introduce and fit a combined CCN-EAM (a 'VAM') to flanker-task-like data. They show that the VAM can fit mean RTs and accuracies as well as the congruency effect that is present in the data, and subsequently analyze the VAM in terms of where in the network congruency effects arise.

I have mixed feelings about this manuscript, as I appreciate the innovative efforts to combine CNNs with EAMs in a new class of cognitive models, while also having some reservations from an EAM perspective. The idea of combining these approaches has great potential, and I'm excited to see where this research will lead. However, I do have some concerns about the quality of fit between the behavioral data and the model. Specifically, the RT distributions, delta plots, and conditional accuracy function don't appear to be well-matched by the VAM. The conflict effects on behavioral data are well-established and typically considered crucial to understanding the underlying cognitive process. Unfortunately, it seems that these parts of the data don't fit well with the proposed model.

This disparity is not entirely surprising. The EAM literature suggests that LBA models might not be suitable for conflict tasks, and the presented results seem to confirm this concern. Conflict EAMs, including the DMC (e.g., Ulrich et al., 2015; Evans & Servant, 2022; Lee & Sewell 2024), propose dynamic drift rates with a fast automatic process that is gradually withdrawn from evidence accumulation over time. This approach results in congruency effects arising from temporal dynamics, not spatial representations.
In contrast, the VAM imposes static drift rates in the LBA model, leading to an effect between drift rates that translates to changes in representations. However, this account does not adequately explain the behavioral data, and the proposed representational geometry explanation is therefore limited.

My concerns are addressed in the revised manuscript, but I struggle to understand why the authors distinguish between explaining mean effects across individuals and congruency effects within individuals. These concepts seem related, and issues at the individual level could propagate to the group mean. Furthermore, I find it challenging to accept that dynamics merely act 'in concert' with the orthogonalization mechanism, as it seems possible that an account that uses a time-varying EAM may not require any orthogonalization mechanism in the first place. The orthogonalization mechanism might have arisen because the model does not have the possibility to account for the conflict effect from temporal effects, instead of spatial effects. I could envision a CNN-DMC in which conflict effects arise only at the level of the choice model (e.g., as a time-varying filter that changes which information is read out from the visual system, rather than due to changes in the representations in the visual system itself). This possibility should be acknowledged in the paper, and it would be interesting to discuss how such an account would be tested.

While I appreciate the technological advancement presented in this paper, my concerns are not about implementation details but rather about the choice of models and their consequences. I believe that a more in-depth exploration of which conclusions can be drawn, and which model comparisons would be required to reach a final conclusion.

https://doi.org/10.7554/eLife.98351.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Public Reviews:***
>
> ***Reviewer #1 (Public Review):***
>
> *Summary:*
>
> *This paper introduces a new approach to modeling human behavioral responses using image-computable models. They create a model (VAM) that is a combination of a standard CNN coupled with a standard evidence accumulation model (EAM). The combined model is then trained directly on image-level data using human behavioral responses. This approach is original and can have wide applicability. However, many of the specific findings reported are less compelling.*
>
> *Strengths:*
>
> *(1) The manuscript presents an original approach to fitting an image-computable model to human behavioral data. This type of approach is sorely needed in the field.*
>
> *(2) The analyses are very technically sophisticated.*
>
> *(3) The behavioral data are large both in terms of sample size (N=75) and in terms of trials per subject.*
>
> *Weaknesses:*
>
> *Major*
>
> *(1) The manuscript appears to suggest that it is the first to combine CNNs with evidence accumulation models (EAMs). However, this was done in a 2022 preprint*
>
> *(https://www.biorxiv.org/content/10.1101/2022.08.23.505015v1) that introduced a network called RTNet. This preprint is cited here, but never really discussed. Further, the two unique features of the current approach discussed in lines 55-60 are both present to some extent in RTNet. Given the strong conceptual similarity in approach, it seems that a detailed discussion of similarities and differences (of which there are many) should feature in the Introduction.*

Thanks for pointing this out—we agree that the novel contributions of our model (the VAM) with respect to prior related models (including RTNet) should be clarified, and have revised the Introduction accordingly. We include the following clarifications in the Introduction:

"The key feature of the VAM that distinguishes it from prior models is that the CNN and EAM parameters are jointly fitted to the RT, choice, and visual stimulus data from individual participants in a unified Bayesian framework. Thus, both the visual representations learned by the CNN and the EAM parameters are directly constrained by behavioral data. In contrast, prior models first optimize the CNN to perform the behavioral task, then separately fit a minimal set of high-level CNN parameters [RTNet, Rafiei et al., 2024] and/or the EAM parameters to behavioral data [Annis et al., 2021; Holmes et al., 2020; Trueblood et al., 2021]. As we will show, fitting the CNN with human data—rather than optimizing the model to perform a task—has significant consequences for the representations learned by the model."

E.g. in the case of RTNet, the variability of the Bayesian CNN weight distribution, the decision threshold, and the magnitude of the noise added to the images are adjusted to match the average human accuracy (separately for each task condition). RTNet is an interesting and useful model that we believe has complementary strengths to our own work.

Since there are several other existing models in addition to the VAM and RTNet that use CNNs to generate RTs or RT proxies (by our count, at least six that we cite earlier in the Introduction), we felt it was inappropriate to preferentially include a detailed comparison of the VAM and RTNet beyond the passage quoted above.

> *(2) In the approach here, a given stimulus is always processed in the same way through the core CNN to produce activations v_k. These v_k's are then corrupted by Gaussian noise to produce drift rates d_k, which can differ from trial to trial even for the same stimulus. In other words, the assumption built into VAM appears to be that the drift rate variability stems entirely from post-sensory (decisional) noise. In contrast, the typical interpretation of EAMs is that the variability in drift rates is sensory. This is also the assumption built into RTNet where the core CNN produces noisy evidence. Can the authors comment on the plausibility of VAM's assumption that the noise is post-sensory?*

In our view, the VAM is compatible with a model in which the drift rate variability for a given stimulus is due to sensory noise, since we do not specify the origin of the Gaussian noise added to the drift rates. As the reviewer notes, the CNN component of the VAM processes a given stimulus deterministically, yielding the mean drift rates. This does not preclude us from imagining an additional (unmodeled) sensory process that adds variability to the drift rates. The VAM simply represents this and other hypothetical sources of variability as additive Gaussian noise. We agree however that it is worthwhile to think about the origin of the drift rate variability, though it is not a focus of our work.

> *(3) Figure 2 plots how well VAM explains different behavioral features. It would be very useful if the authors could also fit simple EAMs to the data to clarify which of these features are explainable by EAMs only and which are not.*

In our view, fitting simple EAMs to the data would not be especially informative and poses a number of challenges for the particular task we study (LIM) that are neatly avoided by using the VAM. In particular, as we show in Figure 2, the stimuli vary along several dimensions that all appear to influence behavior: horizontal position, vertical position, layout, target direction, and flanker direction. Since the VAM is stimulus-computable, fitting the VAM automatically discovers how all of these stimulus features influence behavior (via their effect on the drift rates outputted by the CNN). In contrast, fitting a simple EAM (e.g. the LBA model) necessitates choosing a particular parameterization that specifies the relationship between all of the stimulus features and the EAM model parameters. This raises a number of practical questions. For example, should we attempt to fit a separate EAM for each stimulus feature, or model all stimulus features simultaneously?

Moreover, while we could in principle navigate these issues and fit simple EAMs to the data, we do not intend to claim that simple EAMs fail to explain the relationship between stimulus features and behavior as well as the VAM. Rather, the key strength of the VAM relative to simple EAMs is that it includes a detailed and biologically plausible model of human vision. The majority of the paper capitalizes on this strength by showing how behavioral effects of interest (namely congruency effects) can be explained in terms of the VAM's visual representations.

> *(4) VAM is tested in two different ways behaviorally. First, it is tested to what extent it captures individual differences (Figure 2B-E). Second, it is tested to what extent it captures average subject data (Figure 2F-J). It wasn't clear to me why for some metrics only individual differences are examined and for other metrics only average human data is examined. I think that it will be much more informative if separate figures examine average human data and individual difference data. I think that it's especially important to clarify whether VAM can capture individual differences for the quantities plotted in Figures 2F-J.*

We would like to clarify that Fig. 2J in fact already shows how well the VAM captures individual differences for the average subject data shown in Fig. 2H (stimulus layout) and Fig. 2I (stimulus position). For a given participant and stimulus feature, we calculated the Pearson's r between model/participant mean RTs across each stimulus feature value. Fig. 2J shows the distribution of these Pearson's r values across all participants for stimulus layout and horizontal/vertical position.

Fig. 2G also already shows how well the VAM captures individual differences in behavior. Specifically, this panel shows individual differences in mean RT attributable to differences in age. For Fig. 2F, which shows how the model drift rates differ on congruent vs. incongruent trials, there is no sensible way to compare the models to the participants at any level of analysis (since the participants do not have drift rates).

> *(5) The authors look inside VAM and perform many exploratory analyses. I found many of these difficult to follow since there was little guidance about why each analysis was conducted. This also made it difficult to assess the likelihood that any given result is robust and replicable. More importantly, it was unclear which results are hypothesized to depend on the VAM architecture and training, and which results would be expected in performance-optimized CNNs. The authors train and examine performance-optimized CNNs later, but it would be useful to compare those results to the VAM results immediately when each VAM result is first introduced.*

Thanks for pointing this out—we apologize for any confusion caused by our presentation of the CNN analyses. We have added in additional motivating statements, methodological clarifications, and relevant references to our Results, particularly for Figure 3 in which we first introduce the analyses of the CNN representations/activity. In general, each analysis is prefaced by a guiding question or specific rationale, e.g. "How do the models' visual representations enable target selectivity for stimuli that vary along several irrelevant dimensions?" We also provide numerous references in which these analysis techniques have been used to address similar questions in CNNs or the primate visual cortex.

We chose to maintain the current organization of our results in which the comparison between the VAM and the task-optimized models are presented in a separate figure. We felt that including analyses of both the VAM and task-optimized models in the initial analyses of the CNN representations would be overwhelming for many readers. As the reviewer acknowledges, some readers may already find these results challenging to follow.

> *(6) The authors don't examine how the task-optimized models would produce RTs. They say in lines 371-2 that they "could not examine the RT congruency effect since the task-optimized models do not generate RTs." CNNs alone don't generate RTs, but RTs can easily be generated from them using the same EAM add-on that is part of VAM. Given that the CNNs are already trained, I can't see a reason why the authors can't train EAMs*

> *on top of the already trained CNNs and generate RTs, so these can provide a better comparison to VAM.*

We appreciate this suggestion, but we judge the suggestion to “train EAMs on top of the already trained CNNs and generate RTs” to be a significant expansion of the scope of the paper with multiple possible roads forward. In particular, one must specify how the outputs of the task-optimized CNN (logits for each possible response) relate to drift rates, and there is no widely-accepted or standard way to do this. Previously proposed methods include transforming representation distances in the last layer to drift rates (https://doi.org/10.1037/xlm0000968), fitting additional subject-specific parameters that map the logits to drift rates

(https://doi.org/10.1007/s42113-019-00042-1), or using the softmax-scored model outputs as drift rates directly (https://doi.org/10.1038/s41562-024-01914-8), though in the latter case the RTs are not on the same scale as human data. In our view, evaluating these different methods is beyond the scope of this paper. An advantage of the VAM is that one does not have to fit two separate models (a CNN and a EAM) to generate RTs.

Nonetheless, we agree that it would be informative to examine something like RTs in the task-optimized models. Our revised Results section now includes an analysis of the confidence of the task-optimized models’ decisions, which we use a proxy for RTs:

“Since the task-optimized models do not generate RTs, it is not possible to directly measure RT congruency effects in these models without making additional assumptions about how the CNN's classification decisions relate to RTs. However, as a coarse proxy for RT, we can examine the confidence of the CNN's decisions, defined as the softmax-scored logit (probability) of the most probable direction in the final CNN layer. This choice of RT proxy is motivated by some prior studies that have combined CNNs with EAMs [Annis et al., 2021; Holmes et al., 2020; Trueblood et al., 2021]. These studies explicitly or implicitly derive a measure of decision confidence from the activity of the last CNN layer. The confidence measure is then mapped to the EAM drift rates, such that greater decision confidence generally corresponds to higher drift rates (and therefore shorter RTs).

We calculated the average confidence of each task-optimized CNN separately for congruent vs. incongruent trials. On average, the task-optimized models showed higher confidence on congruent vs. incongruent trials ($W = 21.0$, $p < 1e-3$, Wilcoxon signed-rank test; Cohen's $d = 0.99$; $n = 75$ models). These analyses therefore provide some evidence that task-optimized CNNs have the capacity to exhibit congruency effects, though an explicit comparison of the magnitude of these effects with human data requires additional modeling assumptions (e.g., fitting a separate EAM).”

> *(7) The Discussion felt very long and mostly a summary of the Results. I also couldn't shake the feeling that it had many just-so stories related to the variety of findings reported. I think that the section should be condensed and the authors should be clearer about which explanations are speculations and which are air-tight arguments based on the data.*

We have shortened the Discussion modestly and we have added in some clarifying language to help clarify which arguments are more speculative vs. directly supported by our data.

Specifically, we added in the phrase “we speculate that…” for two suggestions in the Discussion (paragraphs 3 and 5), and we ensured that any other more speculative suggestions contain such clarifying language. We have also added in subheadings in the Discussion to help readers navigate this section.

> *(8) In one of the control analyses, the authors train different VAMs on each RT quantile. I don't understand how it can be claimed that this approach can serve as a model of an*

> *individual's sensory processing. Which of the 5 sets of weights (5 VAMs) captures a given subject's visual processing? Are the authors saying that the visual system of a given subject changes based on the expected RT for a stimulus? I feel like I'm missing something about how the authors think about these results.*

We agree that these particular analyses may cause confusion and have removed them from our revised manuscript.

> ***Reviewer #2 (Public Review):***
>
> *In an image-computable model of speeded decision-making, the authors introduce and fit a combined CCN-EAM (a 'VAM') to flanker-task-like data. They show that the VAM can fit mean RTs and accuracies as well as the congruency effect that is present in the data, and subsequently analyze the VAM in terms of where in the network congruency effects arise.*
>
> *Overall, combining DNNs and EAMs appears to be a promising avenue to seriously model the visual system in decision-making tasks compared to the current practice in EAMs. Some variants have been proposed or used before (e.g., doi.org/10.1016/j.neuroimage.2017.12.078 , doi.org/10.1007/s42113-019-00042-1), but always in the context of using task-trained models, rather than models trained on behavioral data. However, I was surprised to read that the authors developed their model in the context of a conflict task, rather than a simpler perceptual decision-making task. Conflict effects in human behavior are particularly complex, and thereby, the authors set a high goal for themselves in terms of the to-be-explained human behavior. Unfortunately, the proposed VAM does not appear to provide a great account of conflict effects that are considered fundamental features of human behavior, like the shape of response time distributions, and specifically, delta plots (doi.org/10.1037/0096-1523.20.4.731). The authors argue that it is beyond the scope of the presented paper to analyze delta plots, but as these are central to studies of human conflict behavior, models that aim to explain conflict behavior will need to be able to fit and explain delta plots.*
>
> *Theories on conflict often suggest that negative/positive-trending delta plots arise through the relative timing of response activation related to relevant and irrelevant information.*
>
> *Accumulation for relevant and irrelevant information would, as a result, either start at different points in time or the rates vary over time. The current VAM, as a feedforward neural network model, does not appear to be able to capture such effects, and perhaps fundamentally not so: accumulation for each choice option is forced to start at the same time, and rates are a static output of the CNN.*
>
> *The proposed solution of fitting five separate VAMs (one for each of five RT quantiles) is not satisfactory: it does not explain how delta plots result from the model, for the same reason that fitting five evidence accumulation models (one per RT quantile) does not explain how response time distributions arise. If, for example, one would want to make a prediction about someone's response time and choice based on a given stimulus, one would first have to decide which of the five VAMs to use, which is circular. But more importantly, this way of fitting multiple models does not explain the latent mechanism that underlies the shape of the delta plots.*
>
> *As such, the extensive analyses on the VAM layers and the resulting conclusions that conflict effects arise due to changing representations across layers (e.g., "the selection of task-relevant information occurs through the orthogonalization of relevant and irrelevant representations") - while inspiring, they remain hard to weigh, as they are*

> *contingent on the assumption that the VAM can capture human behavior in the conflict task, which it struggles with. That said, the promise of combining CNNs and EAMs is clearly there. A way forward could be to either adjust the proposed model so that it can explain delta plots, which would potentially require temporal dynamics and time-varying evidence accumulation rates, or perhaps to start simpler and combine CCNs-EAMs that are able to fit more standard perceptual decision-making tasks without conflict effects.*

We thank the reviewer for their thoughtful comments on our work. However, we note that the

VAM does in fact capture the positive-trending RT delta plot observed in the participant data (Fig. S4A), though the intercepts for models/participants differ somewhat. On the other hand, the conditional accuracy functions (Fig. S4B) reveal a more pronounced difference between model and participant behavior. As the reviewer points out, capturing these effects is likely to require a model that can produce time-varying drift rates, whereas our model produces a fixed drift rate for a given stimulus. We also agree that fitting a separate VAM to each RT quantile is not a satisfactory means of addressing this limitation and have removed these analyses from our revised manuscript.

However, while we agree that accurately capturing these dynamic effects is a laudable goal, it is in our view also worthwhile to consider explanations for the mean behavioral effect (i.e. the accuracy congruency effect), which can occur independently of any consideration of dynamics. One of our main findings is that across-model variability in accuracy congruency effects is better attributed to variation in representation geometry (target/flanker subspace alignment) vs.

variation in the degree of flanker suppression. This finding does not require any consideration of dynamics to be valid at the level of explanation we pursue (across-user variability in congruency effects), but also does not preclude additional dynamic processes that could give rise to more specific error patterns. Our revised discussion now includes a section where we summarize and elaborate on these ideas:

“It is not difficult to imagine how the orthogonalization mechanism described above, which explains variability in accuracy congruency effects across individuals, could act in concert with other dynamic processes that explain variability in congruency effects within individuals (e.g., as a function of RT). In general, any process that dynamically gates the influence of irrelevant sensory information on behavioral outputs could accomplish this, for example ramping inhibition of incorrect response activation [https://doi.org/10.3389/fnhum.2010.00222], a shrinking attention spotlight [https://doi.org/10.1016/j.cogpsych.2011.08.001], or dynamics in neural population-level geometry [https://doi.org/10.1038/nn.3643]. To pursue these ideas, future work may aim to incorporate dynamics into the visual component and decision component of the VAM with recurrent CNNs [https://doi.org/10.48550/arXiv.1807.00053, https://doi.org/10.48550/arXiv.2306.11582] and the task-DyVA model [https://doi.org/10.1038/s41562-022-01510-8], respectively.”

> ***Reviewer #3 (Public Review):***
>
> *Summary:*
>
> *In this article, the authors combine a well-established choice-response time (RT) model (the Linear Ballistic Accumulator) with a CNN model of visual processing to model image-based decisions (referred to as the Visual Accumulator Model - VAM). While this is not the first effort to combine these modeling frameworks, it uses this combination of approaches uniquely.*

> *Specifically, the authors attempt to better understand the structure of human information representations by fitting this model to behavioral (choice-RT) data from a classic flanker task. This objective is made possible by using a very large (by psychological modeling standards) industry data set to jointly fit both components of this VAM model to individual-level data. Using this approach, they illustrate (among other results) (1) how the interaction between target and flanker representations influence the presence and strength of congruency effects, (2) how the structure of representations changes (distributed versus more localized) with depth in the CNN model component, and (3) how different model training paradigms change the nature of information representations. This work contributes to the ML literature by demonstrating the value of training models with richer behavioral data. It also contributes to cognitive science by demonstrating how ML approaches can be integrated into cognitive modeling. Finally, it contributes to the literature on conflict modeling by illustrating how information representations may lead to some of the classic effects observed in this area of research.*
>
> *Strengths:*
>
> *(1) The data set used for this analysis is unique and is made publicly available as part of this article. Specifically, they have access to data for 75 participants with >25,000 trials per participant. This scale of data/individual is unusual and is the foundation on which this research rests.*
>
> *(2) This is the first time, to my knowledge, that a model combining a CNN with a choice-RT model has been jointly fit to choice-RT data at the level of individual people. This type of model combination has been used before but in a more restricted context. This joint fitting, and in particular, learning a CNN through the choice-RT modeling framework, allows the authors to probe the structure of human information representations learned directly from behavioral data.*
>
> *(3) The analysis approaches used in this article are state-of-the-art. The training of these models is straightforward given the data available. The interesting part of this article (opinion of course) is the way in which they probe what CNN has learned once trained. I find their analysis of how distractor and target information interfere with each other particularly compelling as well as their demonstration that training on behavioral data changes the structure of information representations when compared to training models on standard task-optimized data.*
>
> *Weaknesses:*
>
> *(1) Just as the data in this article is a major strength, it is also a weakness. This type of modeling would be difficult, if not impossible to do with standard laboratory data. I don't know what the data floor would be, but collecting tens of thousands of decisions for a single person is impractical in most contexts. Thus this type of work may live in the realm of industry. I do want to re-iterate that the data for this study was made publicly available though!*

We suspect (but have not systematically tested) that the VAMs can be fitted with substantially less data. We use data augmentation techniques (various randomized image transformations) during training to improve the generalization capabilities of the VAMs, and these methods are likely to be particularly important when training on smaller datasets. One could consider increasing the amount of image data augmentation when working with smaller datasets, or pursuing other forms of data augmentation like resampling from estimated RT distributions (see https://doi.org/10.1038/s41562-022-01510-8 for an example of this). In general, we don’t think that prospective users of our approach should be discouraged if they have only a few hundred trials per subject (or less) - it’s worth trying!

> *(2) While this article uses choice-RT data it doesn't fully leverage the richness of the RT data itself. As the authors point out, this modeling framework, the LBA component in particular, does not account for some of the more nuanced but well-established RT effects in this data. This is not a big concern given the already nice contributions of this article and it leads to an opportunity for ongoing investigation.*

We agree that fully capturing the more nuanced behavioral effects you mention (e.g. RT delta plots and conditional accuracy functions) is a worthwhile goal for future research—see our response to Reviewer #2 for a more detailed discussion. ----------

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations For The Authors):***
>
> *(1) The phrase in the Abstract "convolutional neural network models of visual processing and traditional EAMs are jointly fitted" made me initially believe that the two models were fitted independently. You may want to re-word to clarify.*

We think that the phrase “jointly fitted” already makes it clear that both the CNN and EAM parameters are estimated simultaneously, in agreement with how this term is usually used. But we have nonetheless appended some additional clarifying language to that sentence (“in a unified Bayesian framework”).

> *(2) Lines 27-28: EAMs "are the most successful and widely-used computational models of decision-making." This is only true for the specific type of decision-making examined here, namely joint modeling of choice and response times. Signal detection theory is arguably more widely-used when response times are not modeled.*

Thanks for pointing this out - we have revised the referenced sentence accordingly.

> *(3) Could the authors clarify what is plotted in Figure 2F?*

Fig. 2F shows the drift rates for the target, flanker, and “other” (non-target/non-flanker) accumulators averaged over trials and models for congruent vs. incongruent trials. In case this was a source of confusion, we do not show the value of the flanker drift rates on congruent trials because the flanker and target accumulators are identical (i.e. the flanker/congruent drift rates are equivalent to the target/congruent drift rates).

> *(4) Lines 214-7: "The observation that single-unit information for target direction decreased between the fourth and final convolutional layers while population-level decoding remained high is especially noteworthy in that it implies a transition from representing target direction with specialized "target neurons" to a more distributed, ensemble-level code." Can the authors clarify why this is the only reasonable explanation for these results? It seems like many other explanations could be construed.*

We have added additional clarification to this section and now use more tentative language:

“The observation that single-unit information for target direction decreased between the fourth and final convolutional layers indicates that the units become progressively less selective for particular target directions. Since population-level decoding remained high in these layers, this suggests a transition from representing target direction with specialized "target neurons" to a more distributed, ensemble-level code.”

> *(5) Lines 372-376: "Thus, simply training the model to perform the task is not sufficient to reproduce a behavioral phenomenon widely-observed in conflict tasks. This challenges a core (but often implicit) assumption of the task-optimized training paradigm, namely that to do a task well, a training model will result in model representations that are similar to those employed by humans." While I agree with the general sentiment, I feel that its application here is strange. Unless I'm missing something, in the context of the preceding sentence, the authors seem to be saying that researchers in the field expect that CNNs can produce a behavioral phenomenon (RTs) that is completely outside of their design and training. I don't think that anyone actually expects that.*

We moved the discussion/analyses of RTs to the next paragraph. It should now be clear that this statement refers specifically to the absence of an accuracy congruency effect in the task-optimized models.

> *(6) Lines 387-389: "As a result, the VAMs may learn richer representations of the stimuli, since a variety of stimulus features-layout, stimulus position, flanker direction-influence behavior (Figure 2)." That is certainly true of tasks like this one where an optimal model would only focus on a tiny part of the image, whereas humans are distracted by many features. I'm not sure that this distractibility is the same as "richer representations". When CNNs classify images based on the background, would the authors claim that they have richer representations than humans?*

We agree that “richer” may not be the best way to characterize these representations, and have changed it to “more complex”.

> *(7) Is it possible that drift rate d_k for each response happens to be negative on a given trial? If so, how is the decision given on such trials (since presumably none of the accumulators will ever reach the boundary)?*

It is indeed possible for all of the drift rates to be negative, though we found that this occurred for a vanishingly small number of trials (mean ± s.e.m. percent trials/model: 0.080 ± 0.011%, n = 75 models), as reported in the Methods. These trials were excluded from analyses.

> *(8) Can the authors comment on how they chose the CNN architecture and whether they expect that different architectures will produce similar results?*

Before establishing the seven-layer CNN architecture used throughout the paper, we conducted some preliminary experiments using other architectures that differed primarily in the number of CNN layers. We found that models with significantly fewer than seven layers typically failed to reach human-level accuracy on the task while larger models achieved human-level accuracy but (unsurprisingly) took longer to train.

> ***Reviewer #3 (Recommendations For The Authors)****:*
>
> *- In the introduction to this paper (particularly the paragraph beginning in line 33), the authors note that EAMs have typically been used in simplified settings and that they do not provide a means to account for how people extract information from naturalistic stimuli. While I agree with this, the idea of connecting CNNs of visual processing with EAMs for a joint modeling framework has been done. I recommend looking at and referencing these two articles as well as adjusting the tenor of this part of an introduction to better reflect the current state of the literature. For full disclosure, I am one of the authors on these articles. https://link.springer.com/article/10.1007/s42113-019-00042-1 https://www.sciencedirect.com/science/article/abs/pii/S0010027721001323*

We agree—thanks for pointing this out. The revised Introduction now discusses prior related models in more detail (including those referenced above) and better clarifies the novel contributions of our model. We specifically highlight that a novel contribution of the VAM is that “the CNN and EAM parameters are jointly fitted to the RT, choice, and visual stimulus data from individual participants in a unified Bayesian framework.”

> *- The statement in lines 56-58 implies that this is the first article to glue CNNs together with EAMs. I would edit this accordingly based on the prior comment here and references provided. I will note that the second feature of the approach in this paper is still novel and really nice, namely the fact that the CNN and the EAM are jointly fitted. In the aforementioned references, the CNN is trained on the image set, and individual level Bayesian estimation was only applied to the EAM. Thus, it may be useful to highlight the joint estimation aspect of this investigation as well as how the uniqueness of the data available makes it possible.*

Agreed—see above.

> *- Figure 3c and associated text. I understand the MI analysis you are performing here, however it is difficult to interpret as it stands. In the figure, what does a MI of 0.1 mean?? Can you give some context to that scale? I do find the interpretation of the hunchback shape in lines 210-222 to be somewhat of a stretch. The discussion that precedes (lines 199-209) this is clear and convincing. Can this discussion be strengthened more? And more interpretability of Figure 3c would be helpful; entropic scales can be hard to interpret without some context or scale associated.*

The MI analyses in Fig. 3C (and also Figs. 4C and 6E) show normalized MI, in which the raw MI has been divided by the entropy of the stimulus feature distribution. This normalization facilitates comparing the MI for different stimulus features, which is relevant for Figs. 4C and 6E. The normalized MI has a possible range of [0, 1], where 1 indicates perfect correlation between the two variables and 0 indicates complete independence. We now note in the legend of these figures that the possible normalized MI range is [0, 1], which should help with interpreting these values. Our revised results section for Fig. 3C now also includes some additional remarks on our interpretation of the hunchback shape of the MI.

> *- Lines 244-248 and the analyses in Figure 3 suggest a change in the behavior of the CNN around layer 4. This is just a musing, but what would happen if you just used a 4 layer CNN, or even a 3 layer? This is not just a methods question. Your analysis suggests a transition from localized to distributed information representation. Right now, the EAM only sees the output of the distributed representation. What if it saw the results the more local representations from early layers? Of course, a shallower network may just form the distributed representations earlier, but it would interesting if there were a way to tease out not just the presence of distributed vs local representations, but the utility of those to the EAM.*

Thanks for this interesting suggestion. We did do some preliminary experiments in models with fewer layers, though we only examined the outputs of these models and did not assess their representations. We found that models with 3–5 layers generally failed to achieve human-level accuracy on the task. In principle, one could relate this observation to the representations of these models as a means of assessing the relative utility of distributed/local representations. However, there are confounding factors that one would ideally control for in order to compare models with different numbers of layers in this fashion (namely, the number of parameters).

> *- Section Line 359 (Task optimized models) - It would be helpful to clarify here what these task-optimized models are being trained to do. As I understand it, they are being trained to directly predict the target direction. But are you asking them to learn to predict the true target direction? Or are you training them to predict what each individual responds? I think it is the second (since you have 75 of these), but it's not clear. I looked at the methods and still couldn't get a clear description of this. Also, are you just stripping the LBA off of the end of the CNN and then essentially putting a softmax in its place? If so, it would be helpful to say so.*

The task-optimized models were actually trained to output the true target direction in each stimulus, rather than trained to match the decisions of the human participants. We trained 75 such models since we wanted to use exactly the same stimuli as were used to train each VAM. The task-optimized CNNs were identical to those used in the VAMs, except that the outputs of the last layer were converted to softmax-scored probabilities for each direction rather than drift rates. The Results and Methods section now included additional commentary that clarifies these points.

> *- Line 373-376: This statement is pretty well established at this point in the similarity judgement literature. I recommend looking at and referencing https://onlinelibrary.wiley.com/doi/full/10.1111/cogs.13226 https://www.nature.com/articles/s41562-020-00951-3 https://link.springer.com/article/10.1007/s42113-020-00073-z*

Thanks for pointing this out. For reference, the statement in question is “Thus, simply training the model to perform the task is not sufficient to reproduce a behavioral phenomenon widely-observed in conflict tasks. This challenges a core (but often implicit) assumption of the task-optimized training paradigm, namely that training a model to do a task well will result in model representations that are similar to those employed by humans.”

We agree that the first and third reference you mention are relevant, and we now cite them along with some other relevant work. In our view, the second reference you mention is not particularly relevant (that paper introduces a new computational model for similarity judgements that is fit to human data, but does not comment on training models to perform tasks vs. fitting to human data).

> *- Line 387-388: "VAMs may learn richer representations". This is a bit of a philosophical point, but I'll go ahead and mention it. The standard VAM does not necessarily learn "richer" feature representations. Rather, you are asking the VAM and task-optimized models to do different things. As a result, they learn different representations. "Better" or "richer" is in the eye of the beholder. In one view, you could view the VAM performance as sub-par since it exhibits strange artifacts (congruency effects) and the expansion of dimensionality in the VAM representations is merely a side-effect of poor performance. I'm not advocating this view, just playing devils advocate and suggesting a more nuanced discussion of the difference between the VAM and task-optimized models.*

We agree—this is a great point. We have changed this statement to read “the VAMs may learn more complex [rather than richer] representations of the stimuli”.

> *- Lines 567-570: Here you discuss how the LBA backend of the VAM can't account for shrinking spotlight-like RT effects but that fitting models to different RT quantiles helps overcome this. I find this to be one of the weakest points of the paper (the whole process of fitting RT quantiles separately to begin with). This is just a limitation of the RT component of the model. This is a great paper but this is just a limitation inherent in the model. I don't see a need to qualify this limitation and think it would be better to just*

> *point out that this is a limitation of the LBA itself (be more clear that it is the LBA that is the limiting factor here) and that this leaves room for future research. From your last sentence of this paragraph, I agree that recurrent CNNs would be interesting. I will note that RNN choice-RT models are out there (though not with CNNs as part of the model).*

We agree and have revised this section of the Discussion accordingly (see our response to Reviewer #2 for more detail). We also removed the analyses of models trained on separate RT quantiles.

https://doi.org/10.7554/eLife.98351.2.sa0