TAVAE: A VAE WITH ADAPTABLE PRIORS EXPLAINS CONTEXTUAL MODULATION IN THE VISUAL CORTEX

Anonymous authors
Paper under double-blind review

ABSTRACT

The brain interprets visual information through learned regularities, formalized as performing probabilistic inference under a prior. The visual cortex establishes priors for this inference, some of which are from higher level representations as contextual priors and rely on widely documented top-down connections. While evidence supports that priors are acquired for natural images, it remains unclear if similar separate priors can be flexibly acquired for more specific computations, e.g. when learning a task. To investigate this, we built a generative model trained jointly on natural images and on a simple task, and analyzed it along with large-scale recordings from the early visual cortex of mice. For this, we extended the standard VAE formalism to flexibly and data-efficiently acquire a task such that it reuses representations learned in a task-agnostic manner. The resulting Task-Amortized VAE was used to investigate biases when presenting stimuli that violated the trained task statistics. Such mismatches between the learned task statistics and the incoming sensory evidence resulted in multimodal response profiles, which were also observed in the calcium imaging data from mice performing an analogous task. The task-optimized generative model could account for various characteristics of V1 population activity, including within-day updates to the population responses. Our results confirm that flexible task-specific contextual priors can be learned on-demand by the visual system and can be deployed as early as the entry level of the visual cortex.

1 INTRODUCTION

Deep learning models, including discriminative and generative models, have been shown to successfully model neuronal responses in the visual system of the brain Khaligh-Razavi & Kriegeskorte (2014); Yamins & DiCarlo (2016); Lotter et al. (2020); Csikor et al. (2023); Zhuang et al. (2021). These models were assessed through the efficiency of predicting neuronal responses to natural images or natural videos. However, visual cortical responses are not only determined by the stimulus itself but also by non-stimulus attributes, such as the task the visual system is faced with De Lange et al. (2018); Lange & Haefner (2017; 2022). Notably, recent studies have demonstrated strong and systematic biases as early as in the earliest stage of the visual cortex, the V1 Corbo et al. (2022; 2025). Understanding these systematic biases requires that we understand the computational principles behind the changes occuring at the early stages of processing when learning a novel task.

Systematic biases introduced by task learning can be formalized through learning priors: when the animal's visual cortex adapts to a local context (the task) then the visual system relies on learned regularities of the environment and these learned regularities will in turn affect how visual cortical neurons respond to stimuli. Such a probabilistic interpretation has been advocated in the context of adaptation to the natural environment and it implies that the visual cortex relies on knowledge about the statistics of stimuli in order to interpret an incoming stimulus Yuille & Kersten (2006); Fiser et al. (2010). The entry stage to the visual cortex, the V1, learns elementary features of the environment Hubel & Wiesel (1959), suggesting a representation that has a close to linear relationship with the stimulus Olshausen & Field (1996). Learning a covariance matrix over these features would constitute the most basic form of prior. The hierarchy of visual cortical regions permits a prior that can express richer structures than a covariance matrix, as hierarchically higher layers can establish contextual priors for lower layers through top-down connections in the cortical circuitry Lee & Mumford (2003). Indeed, animal experiments established stimulus influence of higher corti-

cal regions on V1 Lee & Nguyen (2001); Chen et al. (2014); Kok et al. (2016); Ziemba et al. (2019) and these influences were shown to be aligned with the contextual priors acquired by a hierarchical generative end-to-end model trained on natural images Csikor et al. (2023). These contextual priors reflect the learned regularities of the natural environment. In contrast to the priors imposed by the regularities of the natural environment, task-related contextual priors reflect more specific regularities that are only characteristic to the actual task, and these regularities can change from task to task. We propose that contextual priors that reflect task structure are responsible for introducing the systematic biases in the responses of neurons in V1 of animals that were exposed to extensive training of a task.

To investigate task-specific contextual priors, we take a generative modeling approach Lee & Mumford (2003). For this, variational autoencoders (VAEs) provide a flexible framework Kingma & Welling (2013); Rezende et al. (2014). To capture contextual priors, a hierarchical variant of the VAE framework can be employed Kingma et al. (2019); Csikor et al. (2023). The ability of a VAE trained on natural images to manage inference when the test statistics shift is not readily apparent. This is because the encoders, leveraging amortized inference, demand complete retraining with the new task dataset. Nonetheless, this approach is neither efficient in terms of data nor biologically realistic, as it could steer away from a crucial representation obtained during the animal's development stage. Instead, we are seeking a way to build upon a learned representation in different contexts. Flexible learning of novel tasks can rely on recruitment of discrete variables that encode the task structure Rao et al. (2019). As our goal is to investigate how a natural image-trained inference is reshaped by contextual priors, we propose to take a different approach. First, train a non-hierarchical version of the VAE on natural images, which learns a latent representation of the natural images through establishing a generative model. Then, based on this generative model, learn a principled extension of the original VAE capable of flexibly learning contextual priors. Ideally, the learned representation should facilitate the learning of the task, thus reducing the amount of data required. As a natural image trained VAE, a V1-inspired VAE is used that learns a complete representation (dimensionality of latent layer, $\boldsymbol{z}$, is similar to that of the observed layer, $\boldsymbol{x}$) Geadah et al. (2024), more specifically, a version that shows more consistent performance in inference Catoni et al. (2024). We then introduce an extension, Task-Amortized VAE (TAVAE), which reuses the amortized posterior of the task-unaware VAE to obtain task-specific posteriors. We use V1 recordings from mice performing a discrimination task to assess the predictions of TAVAE about response biases. For this, we train the TAVAE on an identical task and contrast inference in TAVAE with population responses in mice.

To find signatures of learned contextual priors of a generative model in the biological brain, we rely on data from a discrimination task where mice learn to make decisions based on simple visual stimuli Corbo et al. (2022; 2025). Extensive training ensures that the animals learn the simple regularities of the task. By recording neural activity through calcium imaging in 10 mice across 6 sessions, we have access to 15,027 neurons while the animals perform the task. Crucially, the task is structured so that the learned stimulus distribution is systematically violated during post-training test sessions, enabling investigation of the biases introduced by the task-specific (contextual) prior.

In this paper, we first introduce the theoretical background for TAVAE. Second, we describe the animal experiment along with the specific hierarchical generative model for V1. Third, we investigate the basic properties of a task-trained contextual prior and its consequences on population responses in V1. Fourth, we identify a signature of competing hypotheses, multimodal responses when the stimulus doesn't match the contextual prior, which we identify in population responses. Fifth, we investigate how updating the contextual prior reshapes response biases and show qualitative agreement with the transformation of population responses within an experimental day as a new stimulus is introduced. Finally, we illustrate how the contextual prior and the stimulus likelihood can be inferred from V1 population responses and show that the inferred likelihood resembles the population activity recorded in animals with no exposure to the discrimination task.

2 Theory of Task-Amortized VAE

Given a distribution of images, $p_0(\boldsymbol{x})$, we want to learn about this distribution by learning a latent representation $\boldsymbol{z}$:

$$p_0(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{z}) p_0(\boldsymbol{z})\, d\boldsymbol{z}, \tag{1}$$

In general, inference about latent features, $p(\boldsymbol{z} \mid \boldsymbol{x})$, in such a model is intractable. Variational Autoencoders (VAEs) Kingma & Welling (2013); Rezende et al. (2014) have been proposed to provide an approximate posterior, $q(\boldsymbol{z} \mid \boldsymbol{x})$, relying on a variational approximation. For this, a lower bound to the log empirical distribution is optimized, called the evidence lower bound (ELBO). Through the optimization of the ELBO a pair of models is learned: the likelihood, $p(\boldsymbol{x} \mid \boldsymbol{z})$, termed the generative model, and the (amortized) variational posterior, $q(\boldsymbol{z} \mid \boldsymbol{x})$, termed the recognition model.

We seek to establish a method to perform inference in a computationally efficient way even if the data generating distribution changed from the natural distribution $p_0(\boldsymbol{x})$ to a task distribution $p_T(\boldsymbol{x})$. By default, adapting a VAE to new data requires retraining both the generative and recognition models from scratch by re-optimizing the ELBO. However, in many cases variation in the new dataset that occurs in the task is limited, thus the efficiency of optimizing the original set of parameters is also limited. Instead, we seek a computationally efficient way of adapting to different tasks by reusing the originally learned neural networks. We propose that the latent representation learned by the original VAE is retained as these learned latents can establish a useful feature space for the task. Consequently, we propose that the new generative model retains $p(\boldsymbol{x} \mid \boldsymbol{z})$ in the new context and the change in the stimulus distribution is solely induced by change in the latent prior $p_T(\boldsymbol{z})$:

$$p_T(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{z}) p_T(\boldsymbol{z})\, d\boldsymbol{z}. \tag{2}$$

The task prior, $p_T(\boldsymbol{z})$, can also be thought of as the marginal of higher level latents in a hierarchical model, e.g. a mixture of options, o, and thus $p_T(\boldsymbol{z}) = \sum_{\mathrm{o}} p_T(\boldsymbol{z} \mid \mathrm{o})\, p(\mathrm{o})$. In this generative model, the new posterior is obtained by applying Bayes rule:

$$p_T(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{p_0(\boldsymbol{z} \mid \boldsymbol{x}) p_T(\boldsymbol{z})}{p_0(\boldsymbol{z})} \frac{1}{N(\boldsymbol{x})}, \tag{3}$$

where the normalization constant is

$$N(\boldsymbol{x}) = \frac{p_T(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \int d\boldsymbol{z}\, \frac{p(\boldsymbol{z} \mid \boldsymbol{x}) p_T(\boldsymbol{z})}{p_0(\boldsymbol{z})}. \tag{4}$$

We can approximate the true posterior, $p_0(\boldsymbol{z} \mid \boldsymbol{x})$ in Eq. 3 with the variational posterior, $q(\boldsymbol{z} \mid \boldsymbol{x})$ coming from the trained VAE.

Optimizing the task prior. Eq. 3 highlights that once we know the prior related to the task, one can use the natural variational posterior to obtain the posterior under the prior associated to the task. To achieve this, we derive an optimization objective to determine the task prior itself. We would like to maximize the log-likelihood under the latent prior, $p_T(\boldsymbol{z})$ using observations from the task X_T:

$$L = \sum_{\boldsymbol{x} \in X_T} \log(p_T(\boldsymbol{x})) = \sum_{\boldsymbol{x} \in X_T} \log \int d\boldsymbol{z}\, p(\boldsymbol{x} \mid \boldsymbol{z})\, p_T(\boldsymbol{z}) \tag{5}$$

If we assume that the variational posterior of the natural images are good approximation of the true posterior

$$p(\boldsymbol{x} \mid \boldsymbol{z}) \approx \frac{q(\boldsymbol{z} \mid \boldsymbol{x})\, p_0(\boldsymbol{x})}{p_0(\boldsymbol{z})}, \tag{6}$$

then the log-likelihood, with $p_0(\boldsymbol{x})$ moved outside of the $d\boldsymbol{z}$ integral, can be written as:

$$L = \sum_{\boldsymbol{x} \in X_T} \left[\log p_0(\boldsymbol{x}) + \log \int d\boldsymbol{z}\, \frac{q(\boldsymbol{z} \mid \boldsymbol{x})\, p_T(\boldsymbol{z})}{p_0(\boldsymbol{z})} \right] \tag{7}$$

As the first term does not depend on the task prior, the functional to maximize is:

$$L' = L - L_0 = \sum_{\boldsymbol{x} \in X_T} \log \frac{p_T(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \sum_{\boldsymbol{x} \in X_T} \log \int d\boldsymbol{z}\, \frac{q(\boldsymbol{z}|\boldsymbol{x}) p_T(\boldsymbol{z})}{p_0(\boldsymbol{z})}, \tag{8}$$

where L_0 is the log-likelihood of the natural data.

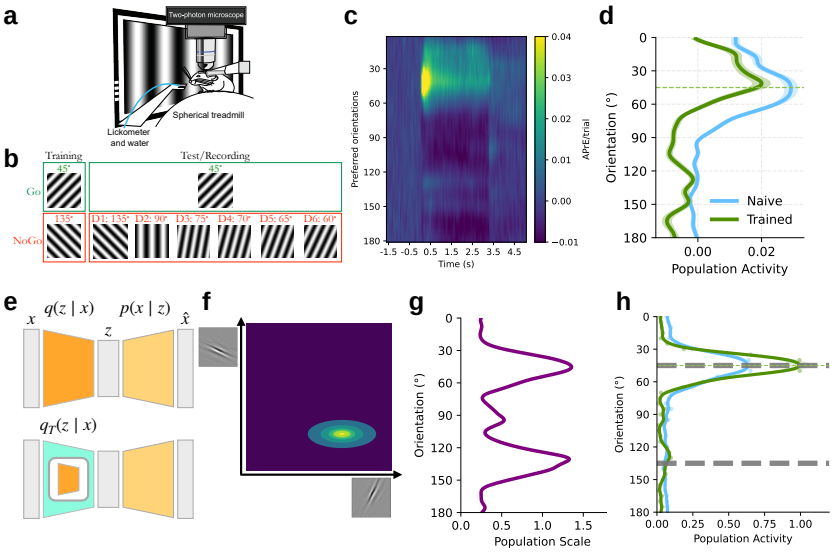

Figure 1: **a**, Experimental setup. **b**, Go-NoGo visual discrimination task stimuli with training and testing schedule. **c**, Population response map of the recorded V1 population on D1 for the Go stimulus. **d** Population response profile averaged over time for trained and untrained (naive) animals. **e** Cartoon of VAE (top) and TAVAE (bottom) models. The TAVAE generative model is retained from the standard VAE whereby the recognition model is transformed. **f**, Illustration of the posterior of two latent dimensions; inset: receptive fields of latents. **g**, Task prior of TAVAE for the discrimination task. Scale of the Laplace prior is shown. **h**, TAVAE responses to the Go stimulus with task and natural priors

3 Experimental orientation discrimination task and its VAE implementation

Experimental paradigm. We tested the predictions of learning a contextual prior in a generative model on neuron population data recorded from the primary visual cortex of mice that were performing a discrimination task Corbo et al. (2025). Briefly, mice performed a visual Go-NoGo task in which animals are required to lick for a 45° or withhold from licking for a 135° moving grating stimulus (Fig. 1a). Animals were first trained on the orientation discrimination task until reaching proficient performance, after which they underwent a six-day testing period accompanied by neural population recordings. During the testing period, the orientation of the Go stimulus remained the same as in training (45°), while the NoGo stimulus was progressively shifted toward the Go orientation, reducing the angular difference to 15° by Day 6 (Fig. 1b). Calcium recordings of excitatory neurons in layer 2/3 of V1 (number of neurons 2007–2675 per session) were performed during the test phase from 10 GCaMP6f or 6s expressing mice in six different conditions. The fractional fluorescence (dF/f) signal obtained from calcium imaging was deconvolved to infer action potential-related events (APrEs), providing a proxy for spiking activity. Neural activity was characterized by the orientation space activity profile (which we will call the population response profile in this paper for simplicity), where neurons are arranged according to the estimated peak of their orientation tuning curve (Fig. 1d), measured in a separate tuning block, and their response intensity is measured by averaging across time. For more details see Corbo et al. (2025). Population response profiles were merged across animals thus reducing recording noise.

Task-Amortized VAE for the discrimination task. We studied the contribution of contextual priors on an end-to-end trained VAE that reflected basic properties of V1. Activations of latent variables, z, of the generative model were assumed to correspond to activations of individual neurons in V1. The generative model assumed that a task prior is learned that reflects the response distribution of latents when task stimuli are presented. This task prior corresponds to a mixture of the two options. TAVAE permits the explicit modeling of the choice the animals can make as binary stochastic

variable, c and then the choice would define conditional priors, $p(\boldsymbol{z} \mid \mathrm{c})$. We opted to model the marginal contextual prior instead. Acquiring the task was modeled by learning the marginal prior, $p_T(\boldsymbol{z})$, according to Eq. 8 (Fig. 1e). To mimic the receptive field properties of neurons (specifically simple cells) of V1, we trained the baseline VAE, $p_0(\boldsymbol{x})$ on natural image patches. These patches were 40-pixel whitened crops from the van Hateren data base Van Hateren & van der Schaaf (1998). The VAE had four characteristic features. 1, Following an earlier VAE study Geadah et al. (2024), the generative model was linear, analogous to independent component accounts of V1 organization Olshausen & Field (1996); 2, The latent space was large dimensional (1799 dimensions) mimicking the complete / overcomplete structure of V1 and ensuring that the linear generative model can accommodate the information in the observations; 3, The prior over latents was Laplace, motivated by efficient coding considerations, and also aligned with earlier accounts Olshausen & Field (1996). In practice, this condition is necessary to obtain a localized, oriented receptive field: a key property of V1 neurons; 4, A scaling latent was introduced, mimicking the Gaussian Scale Mixture (GSM) family of models of natural images Wainwright & Simoncelli (1999). This choice is motivated by the joint statistics of natural images, the superior performance of GSM's to predict V1 responses, and an earlier study which showed that this extended version of VAE ensures more reliable inference, especially in lower contrast regimes Catoni et al. (2024). In summary, the generative model of this baseline model was $p_\varphi(\boldsymbol{x} \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}; \exp(s) \cdot \boldsymbol{A}\boldsymbol{z}, \sigma^2 \boldsymbol{I})$, $p_0(\boldsymbol{z}) = \mathrm{Laplace}(\boldsymbol{z}; 1, 0)$, $p(s) = \mathcal{N}(s; 0, 1)$, where $\exp(s)$ ensures that the scaling is positive. Based on this generative model, we optimize the ELBO:

$$\mathcal{L}(\mathbf{x}, \varphi, \theta, \psi) = -\mathbb{E}_{q_\theta(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\varphi(\boldsymbol{x}|\boldsymbol{z}, s)\right] + D_{KL}(q_\theta(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) + D_{KL}(q_\psi(s|\boldsymbol{x})||p(s)) \quad (9)$$

such that the variational posteriors $q_\theta(\boldsymbol{z} \mid \boldsymbol{x})$ and $q_\psi(\mathrm{s} \mid \boldsymbol{x})$ were modeled as Laplace and Normal distributions, respectively (we will omit the subscript from now on). Since the variable $\boldsymbol{z}$ can be negative we will identify the mean neuronal response to a given image (which measured as APrE in the experiment) as the absolute value of the mean of the posterior $\left|\mathbb{E}_{q_T(z|x)}[z]\right|$, similar to Csikor et al. (2023); Geadah et al. (2024)).

To learn the recognition model for the (discrimination) task, we applied the TAVAE formalism. We approximated the prior with a diagonal covariance matrix. Since the variational posterior (and also the prior) is factorized $q(\boldsymbol{z}, s \mid \boldsymbol{x}) = q(\mathrm{s} \mid \boldsymbol{x}) \cdot q(\boldsymbol{z} \mid \boldsymbol{x})$ on the scale and V1 type latent variables Eq. 3 translate to:

$$q_{\mathrm{T}}(\boldsymbol{z}|\boldsymbol{x}) = \frac{q(\boldsymbol{z}|\boldsymbol{x})\, p_{\mathrm{T}}(\boldsymbol{z})}{p_0(\boldsymbol{z})} \quad (10)$$

where $p_{\mathrm{T}}(\boldsymbol{z})$ was the prior over latents defined by the task. We search for an appropriate task prior as a zero mean Laplace distribution. We choose the mean to be constrained to be zero, since in a typical gratings dataset we expect a symmetry in $\boldsymbol{z}$ around zero:

$$p_T(\boldsymbol{z}) = \mathrm{Laplace}(\boldsymbol{z}; \underline{\sigma}_{\mathrm{T}}, 0) = \prod_{i=1}^{N} \frac{1}{2\sigma_{T,i}} \exp\left(-\frac{|z_i|}{\sigma_{T,i}}\right). \quad (11)$$

We trained the task prior by learning the component of the variances for the diagonals when presenting task stimuli. Thus, the vector $\underline{\sigma}_{\mathrm{T}}$ summarizes the learned parameters (Fig. 1g).

The scales can be determined in a principled way by taking the derivative of the log-likelihood with respect to the scales.

$$0 = \frac{\partial L'}{\partial \sigma_{T,i}} = -\frac{n}{\sigma_{T,i}} + \sum_{\boldsymbol{x} \in X_T} \frac{1}{\sigma_{T,i}^2} \frac{\displaystyle\int dz_i \, \frac{q(z_i \mid \boldsymbol{x})}{p_0(z_i)} |z_i| \exp\left(-\frac{|z_i|}{\sigma_{T,i}}\right)}{\displaystyle\int dz_i \, \frac{q(z_i \mid \boldsymbol{x})}{p_0(z_i)} \exp\left(-\frac{|z_i|}{\sigma_{T,i}}\right)}, \quad (12)$$

where n is the number of images in the task dataset. This can be rearranged to an intuitive form:

$$\sigma_i = \frac{1}{n} \sum_{\boldsymbol{x} \in X_T} \mathbb{E}_{q_{T,\sigma}(z_i|\boldsymbol{x})}\left[|z_i|\right] \quad (13)$$

This is a self-consistency equation, since, of course, the right-hand side also depends on σ_i. We solve this iteratively. We start with the original prior $p_T^{(0)}(\boldsymbol{z}) = p_0(\boldsymbol{z})$. Then after the first iteration, the scales for each dimension will correspond to the scale from the variational posterior, which will be refined in the subsequent iterations. The resulting approximate posterior (Eq. 10) is no longer a Laplace distribution but a piecewise exponential function. Because of this, the necessary integrals can be performed analytically to calculate the various moments, as we demonstrate in Section A.3.

Modeling population response profile using TAVAE. To obtain a population response profile analogous to the one yielded by experiments, we first labeled each z component by orientation preference. For this we synthesized a dataset which contained gratings of different angles and phases. For each component we calculated the response to different angles (averaged over the different phases). By doing this we constructed the orientation tuning curve for each model neuron (similarly as in an actual experiment). We fitted two curve to the tuning curve: a von-Mises function and a constant function. If the former fit was better by a given threshold (we applied a 0.5 threshold in improvement R^2), then we classified that unit as orientation selective with the preferred orientation determined by the best fit (tuning profile for example neurons are shown in Appx. Fig. 5). We found that the overwhelming majority of model neurons (1415 / 1799) were orientation sensitive. After arranging units z_i according to their preferred orientations, we calculated the population response profile by making 5° bins in the model and calculate the average response to the stimulus for those components corresponding to the given bin.

4 RESULTS

Effect of contextual prior for task-congruent stimuli. First, we tested how task-specific training influenced model inference and compared the effects of contextual priors in task-trained versus naïve animals. We focused on the first recording session (D1), in which the test stimuli matched those used during training. To align the model with the experimental paradigm, we synthesized 45° and 135° full-contrast gratings of varying phases (see A.1) to emulate motion, and fit the task prior using Eq.13. The optimization converged after five iterations (Appx. Fig.6). Using the acquired prior, we computed population response profiles for both stimulus orientations (Fig. 2a,b), averaging over all phases. We used a reduced contrast stimulus for testing, in line with the experiment data Corbo et al. (2025).

To understand the effect of a contextual prior, we contrasted the task-optimized TAVAE response profiles with the task-general, standard VAE profiles (Fig. 2a,b). Similarly, we compared experimental response profiles recorded in sessions when the Go and NoGo stimuli were identical with training stimuli (45° and 135°, D1) with response profiles recorded in naive animals. The task-optimized contextual prior resulted in sharpening of the response profile (Fig. 2c). Similar sharpening was evident in mice when analyzing population activity (Fig. 2g).

Another signature of a task-optimized generative model is a significant reduction in the baseline activity, which is consistent across stimulus conditions (Fig. 2a,c). Baseline activity in model V1 neuron responses in the task-agnostic VAE is a result of a broad prior that is flat in orientation space. Intuitively, if the likelihood is not peaking sharply in the orientation space then sampling a posterior that displays residual uncertainty results in small but consistent activity. Upon learning, the selective sharpening of the prior (Fig. 1g) for neurons that have preferred stimulus orientation not matching the stimulus results in a suppressed baseline. Similar systematic baseline activity reduction was also characteristic in mouse recordings, which can be identified through comparing activity levels of neurons not driven by task stimuli (Fig. 2h).

In the task-optimized TAVAE, suppression of baseline activity is not homogeneous: when presenting one task stimulus, suppression does not affect neurons tuned to training stimulus orientations (Fig. 2a). This can be explained by the heterogeneity of the prior: along directions where the prior is wider (the trained orientations), the posterior will also be wider and this uncertainty will result in baseline activity similar to the VAE with the natural prior even though the posterior is centered on zero. When considering experimental recordings, the magnitude of the heterogeneity in baseline suppression is close to residual fluctuations. Although such inhomogeneity can occasionally be experimentally observed (see the bump around the 135° NoGo orientation in Fig. 2e), residual variance prevents a general conclusion.

Systematic biases of the contextual priors for OOD stimulus. In five experimental sessions, animals were exposed to stimuli that systematically deviated from the trained stimulus set. More precisely, the NoGo stimulus gradually converged to the Go stimulus (Fig. 1). Population response profiles from these sessions show drastic deviations from the response predicted merely by the stimulus orientation (Fig. 3a). In particular, multimodal responses are characteristic in multiple con-

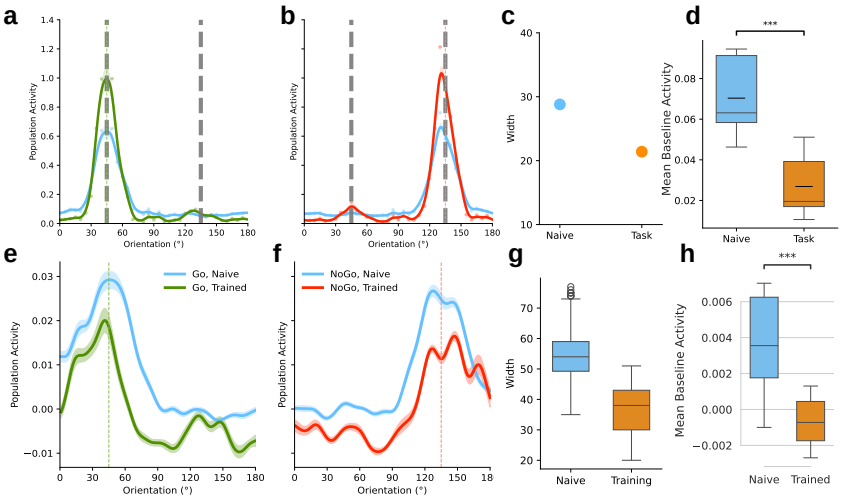

Figure 2: **Effect of learning the discrimination task in the model and in the experiment. a, b**, Response profiles to task stimuli (45° and 135°, gratings, red and green dashed lines, respectively) using natural prior (blue) and the task prior (solid green and red) in the model. Dashed grey lines correspond to the learned contextual prior. **c**, Width of the 45° peak using the natural prior (blue) and task prior (orange). **d**, Baseline reduction in the model. **e, f**, D1 Go and NoGo responses in naive (blue) and trained (green, red) mice. **g**, Width of the Go response aggregated over days for naive (blue) vs. trained animals (orange). **h**, Baseline reduction due to training in the experiment.

ditions (especially in D4 and D5). Contraintuitively, multimodal responses display troughs at the actual stimulus orientation, with peaks flanking the stimulus orientation from both sides (Fig. 3a).

We assumed that these multimodal responses are a consequence of the uncertainty about the source of the stimulus when the contextual prior does not match the observation. Indeed, when simulating experimental population responses across the five conditions (corresponding to experimental days) such that the contextual prior was matched to the stimuli (corresponding to the mice learning a new prior every day), no bimodal profiles were observed (Appx. Fig. 7a). Thus, we assessed TAVAE responses under two more different assumptions. First, we assumed that the contextual prior acquired during the extensive training is retained, i.e. we assumed a prior at 45° and 135° (Appx. Fig. 7b). While showing some effect of bimodality, the response profile matched the recordings poorly. Second, we reasoned that the abrupt change in the NoGo stimulus from 135° to 90° on D2 likely induced an update in the prior. This interpretation is supported by the observation that the false alarm rate on Day 2 is comparable to that on Day 1 (Appx. Fig. 8a), and notably, the false alarm rate decreases across behavioral quintiles within Day 2 (Appx. Fig. 8b). In contrast, from Day 3 onward, the false alarm rate steadily increases (Appx. Fig. 8a). Confirming this assumption, a prior that reflects 45° Go and 90° NoGo stimuli produces these systematic biases, including bimodality for the NoGo stimulus under the conditions used during D2 to D6 (Fig. 3b). Crucially, these biases are such that peak activities do not match either of the components of the contextual prior. Remarkably, the orientaion prior for the Go signal is expected to be consistent across sessions and peaked at 45° and therefore a consistent shift of the mode away from this orientation indicates a systematic bias. From a Bayesian perspective, the systematic shift naturally arises from the combination of an orientation prior and likelihood. This supports how the contextual priors that govern the discrimination task-optimized TAVAE lead to biases congruent with biases in V1 activity in mice adapted to the discrimination task.

Signature of updating the contextual prior. Our results indicate that D1 responses are consistent with a contextual prior peaking at 45° and 135°. Further analysis of D2 through D6 sessions indicated a different contextual prior peaking at 45 and 90°. Taking these findings together, animals seem to shift their priors from the first session (D1) when stimuli are identical with training stimuli to the second session when the NoGo stimulus is radically updated. Behavioral results indicate

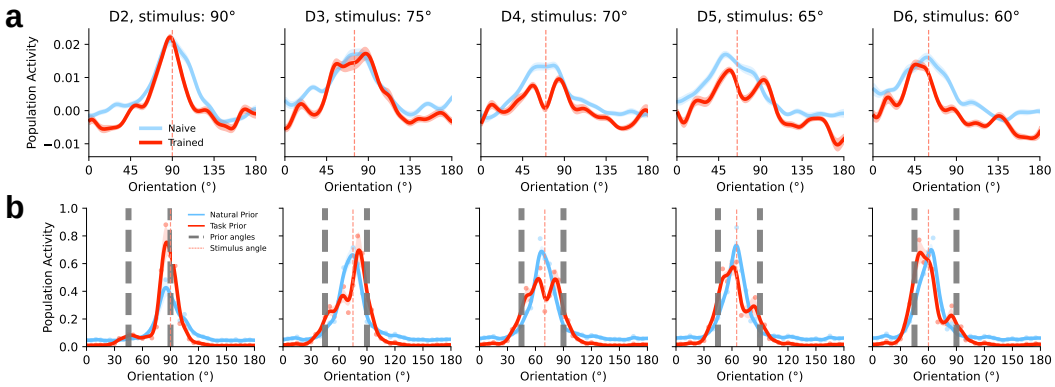

Figure 3: **Systematic biases for mismatched training and test in experiment and model**. **a**, V1 responses over five experimental sessions during which the NoGo stimulus (red) is deviating from the training NoGo stimulus at 135°. Responses from naive animals are shown for reference (blue). **b**, Model responses on the same stimulus orientations as in the experiment across the five sessions. A prior learned for 45° and 90° gratings is assumed (grey dashed lines).

an updating of priors during D2, motivating an analysis of population activity throughout the D2 experimental session.

We model the shifting prior by fitting the task prior to a dataset with gratings characterized by unbalanced orientation classes. Namely, we investigated a dataset that contains 45°, 90°, and 135° gratings with the ratio $1 : \gamma : 1-\gamma$. We obtained contextual priors for $\gamma \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$ and calculated population response profiles to the D2 NoGo stimulus, oriented at 90°.

Closer inspection of high-weighted 135° priors (i.e. $\gamma < 0.5$) reveals three modes: surrounding the central, stimulus-aligned model are two flanking peaks (Fig. 4a). These flanking modes correspond to (shifted) modes contributed by the alternative hypotheses, i.e. that the stimulus is coming from the 45° and 135° components of the prior. A low weight of the 135° stimulus in the contextual prior results in asymmetric flanking modes whereby the mode corresponding to the 45° component is greater than the 135° (Fig. 4a). Population response profiles calculated for the D2 activity of mice exhibit a similar trend. To show this, we stratified D2 trials into five quintiles and analyzed population response profiles separately. Similar to TAVAE, flanking modes are consistently present both early and late in the session (Fig. 4b). Importantly, as expected from theory, the flanking modes are 'attracted' towards the orientation of the NoGo stimulus, i.e. to 90°, in line with the formation of an orientation posterior by combining a contextual prior over orientations with a stimulus-driven likelihood. Further confirming our hypothesis, late trials display asymmetrical flanking modes. This phenomenon is seen both in TAVAE, where decreasing the relative weight of the trained prior leads to a suppressed 135° mode, and in the experiment, where later quintiles of session D2 show a comparable trend (Fig.4c,d). This shift is also reflected behaviorally by a gradual reduction in false alarm rates across quintiles (Appx. Fig. 8b).

We argue that shifts of 45° and 135° modes towards the NoGo stimulus on D2 is a signature of probabilistic inference under uncertainty: a wide likelihood combined with a finite-width prior results in a shifted posterior. We then use this insight to estimate the likelihood and the orientation prior from the position of the modes from the population response profiles (see A.6 for derivation). The inferred likelihood is substantially wider than the orientation prior (Fig. 4e,f), which is consistent with the relatively small shift of the modes away from the originally trained stimulus orientations. Notably, this wide, low-fidelity likelihood resembles the width of the population response profile observed in untrained animals. In untrained animals, where the orientation prior is effectively flat and cannot sharpen the posterior, the population response profile provides a close approximation of the likelihood itself.

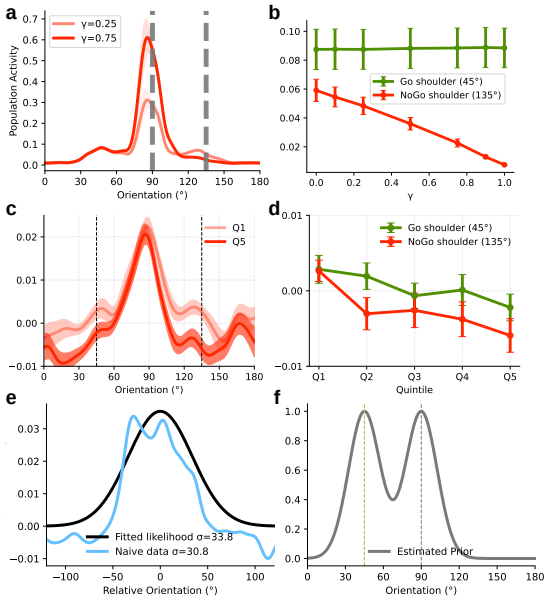

Figure 4: **Signatures of updating a contextual prior**. **a**, TAVAE population response profiles under a mixture contextual prior, where the weight (γ) of the 90° components is increased from 0.25 to 0.75. **b**, Progression of the heights of flanking modes with changing mixture components simulating the progression of the trials across the 90° stimulus session for TAVAE. **c**, Population response profile on D2 when 90° stimulus is introduced, for early (Quintile 1) and late trials (Quintile 5). **d**, as *b* but for experiment D2. **e**, Likelihood inferred from D2 population response profile (black, centered on 0°) and population response profile of naive animals (blue, centered on 0°). **f**, D2 orientation prior inferred from population response profile.

5 Discussion

In this paper, we investigated the biases that emerge from task learning in a generative model and compared them to the biases observed in animals extensively trained on the same task. To study this problem, we developed a variant of VAEs, the TAVAE in which the prior can be flexibly adapted in a computationally efficient way. This permitted the reuse of the representation acquired through exposure to natural image statistics (a useful core representation for the animal). TAVAE could account for a range of basic phenomena, including the sharpening of population responses with learning (without directly changing the receptive fields of neurons) and changes in baseline. Crucially, the model revealed bimodal responses, a hallmark of probabilistic inference under uncertainty, and this bimodality tightly aligned with that observed in V1 when trained priors were not matched with evidence. TAVAE could also capture the updating of the prior when the animal was faced with updated task contingencies.

We chose the task prior as a simple modification of the natural prior, namely by retaining independence among latent variables and tuning only their variances. Interestingly, although the posterior in the high-dimensional latent space was unimodal, this still translated into a multimodal population response profile in the one-dimensional orientation space within a certain parameter range. The TAVAE is capable of accommodating more complex priors that can be applied to more complex tasks and can lead to more specific predictions.

TAVAE permits the explicit reuse of the recognition model when acquiring a new task. Although we do not argue that this algorithmic solution is the one used by the neural circuitry, it is important to consider how the circuitry actually supports these computations. The systematic distortions of neuronal responses observed in the experiment were recorded from layer 2/3 neurons, where bottom-up and top-down signals are integrated. The laminar organization of the cortex might permit more complex computations, for instance layer 4 neurons might actually preserve the original recognition model to be modulated by top-down influences when the feed-forward input reaches layer 2/3. It is a delicate question how a task-related contextual prior is integrated with the contextual priors reflecting the regularities of natural images Csikor et al. (2023). The representational geometry of population responses might give a clue about this Lazar et al. (2024) but this remains an open question.

In principle, the framework described here can also be applied to model contextual modulations in higher cortical areas. More broadly, deep generative models may become an important tool for modeling cortical activity under realistic conditions.

Reproducibility statement

The mathematical derivation needed for the results can be found in the main text 2, 3 and in the appendix A.3. The code used for generating the results can be found in the Supplementary material. In the code, we provided a script that downloads the model weights of the VAE used in the paper.

References

Josefina Catoni, Domonkos Martos, Ferenc Csikor, Enzo Ferrante, Diego H Milone, Balázs Meszéna, Gergő Orbán, and Rodrigo Echeveste. Uncertainty in latent representations of variational autoencoders optimized for visual tasks. *arXiv preprint arXiv:2404.15390*, 2024.

Minggui Chen, Yin Yan, Xiajing Gong, Charles D Gilbert, Hualou Liang, and Wu Li. Incremental integration of global contours through interplay between visual cortical areas. *Neuron*, 82(3): 682–694, 2014.

Julien Corbo, John P McClure, O Batuhan Erkat, and Pierre-Olivier Polack. Dynamic distortion of orientation representation after learning in the mouse primary visual cortex. *Journal of Neuroscience*, 42(21):4311–4325, 2022.

Julien Corbo, O Batuhan Erkat, John McClure Jr, Hussein Khdour, and Pierre-Olivier Polack. Discretized representations in v1 predict suboptimal orientation discrimination. *Nature communications*, 16(1):41, 2025.

Ferenc Csikor, Balázs Meszéna, Katalin Ócsai, and Gergő Orbán. Top-down perceptual inference shaping the activity of early visual cortex. *BioRxiv*, pp. 2023–11, 2023.

Floris P De Lange, Micha Heilbron, and Peter Kok. How do expectations shape perception? *Trends in cognitive sciences*, 22(9):764–779, 2018.

József Fiser, Pietro Berkes, Gergő Orbán, and Máté Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3):119–130, 2010.

Victor Geadah, Gabriel Barello, Daniel Greenidge, Adam S Charles, and Jonathan W Pillow. Sparse-coding variational autoencoders. *Neural computation*, 36(12):2571–2601, 2024.

David H Hubel and Torsten N Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Peter Kok, Lauren J Bains, Tim Van Mourik, David G Norris, and Floris P de Lange. Selective activation of the deep layers of the human primary visual cortex by top-down feedback. *Current Biology*, 26(3):371–376, 2016.

Richard D Lange and Ralf M Haefner. Characterizing and interpreting the influence of internal variables on sensory activity. *Current opinion in neurobiology*, 46:84–89, 2017.

Richard D Lange and Ralf M Haefner. Task-induced neural covariability as a signature of approximate bayesian learning and inference. *PLoS computational biology*, 18(3):e1009557, 2022.

Andreea Lazar, Liane Klein, Johanna Klon-Lipok, Mihály Bányai, Gergő Orbán, and Wolf Singer. Paying attention to natural scenes in area v1. *Iscience*, 27(2), 2024.

Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, 20(7):1434–1448, 2003.

Tai Sing Lee and My Nguyen. Dynamics of subjective contour formation in the early visual cortex. *Proceedings of the National Academy of Sciences*, 98(4):1907–1911, 2001.

William Lotter, Gabriel Kreiman, and David Cox. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature machine intelligence*, 2(4):210–219, 2020.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Dushyant Rao, Francesco Visin, Andrei Rusu, Razvan Pascanu, Yee Whye Teh, and Raia Hadsell. Continual unsupervised representation learning. *Advances in neural information processing systems*, 32, 2019.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

J Hans Van Hateren and Arjen van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1394):359–366, 1998.

Martin J Wainwright and Eero Simoncelli. Scale mixtures of gaussians and the statistics of natural images. *Advances in neural information processing systems*, 12, 1999.

Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006.

Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

Corey M Ziemba, Richard K Perez, Julia Pai, Jenna G Kelly, Luke E Hallum, Christopher Shooner, and J Anthony Movshon. Laminar differences in responses to naturalistic texture in macaque v1 and v2. *Journal of Neuroscience*, 39(49):9748–9756, 2019.

A Appendix

A.1 Gratings dataset

We generated a synthetic dataset consisting of 40×40 grayscale images of sinusoidal gratings. Each image is defined by the function

$$g(X, Y) = C \cdot \sin\Big(2\pi f \cdot \big(X\cos(\theta) + Y\sin(\theta)\big) + \phi\Big),$$

where f represents the spatial frequency (fixed at 3), θ is the orientation angle (measured in radians), ϕ stands for the phase, and C serves as the contrast scaling factor.

The coordinates (X, Y) correspond to pixel locations on a uniform 40×40 grid spanning the interval $[-1, 1]$ in both dimensions. For fitting the scales of the task prior, we selected a contrast factor of 1, while for inputting the test stimuli into the model, we opted for a contrast factor of 0.3 to simulate the reduced contrast used during the test phase of the experiment. The dataset comprised 36 angles spanning from 0° to 180°, along with 50 distinct phase values.

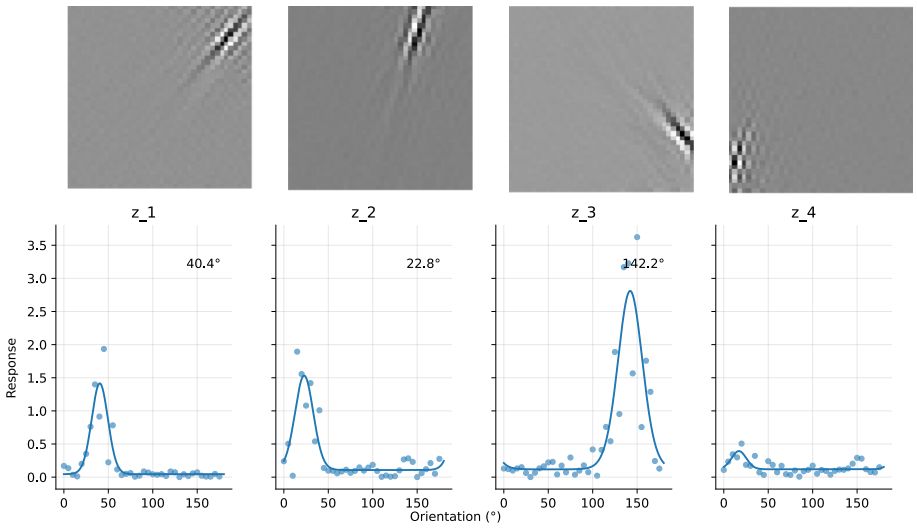

Figure 5: **a**, Receptive fields for example model neurons in EAVAE baseline model **b** Tuning curves and von-misses fits for these neurons. Note that the last neuron is not orientation selective.

A.2 Receptive fields and tuning curves of EVAE

The latent variables z of our baseline VAE model EAVAE have a linear receptive field (determined by the image produced when just one z value is assigned a 1 and all others are set to zero) consist of localized oriented filters (see Fig. 5 a).

Thus, it is reasonable to infer that (some) of these model neurons demonstrate orientation selectivity when gratings at different angles pass through the network, as illustrated by the examples in Fig. 5b

A.3 Calculating moments for products of Laplacian distributions

We are interested in computing the expectation values

$$\mathbb{E}[z_i], \quad \mathbb{E}[|z_i|],$$

under an unnormalized distribution of the form

$$f(z) \;\propto\; q(z;\mu,\sigma)\,\frac{p_T(z;0,\sigma_T)}{p_0(z;0,\mathbf{1})}, \quad z\in\mathbb{R}^N.$$

Here q is a multivariate Laplace with mean vector $\mu\in\mathbb{R}^N$ and scale vector $\sigma\in\mathbb{R}^N$, p_T is a zero-mean Laplace with scale vector $\sigma_T\in\mathbb{R}^N$, and p_0 is the standard zero-mean, unit-scale Laplace.

Because Laplace densities factorize across coordinates, both expectations reduce to one-dimensional problems of the form

$$f_i(z_i) \;\propto\; \exp\Big(-\tfrac{|z_i-\mu_i|}{\sigma_i}-\tfrac{|z_i|}{\sigma_{T,i}}+|z_i|\Big).$$

The integrand is piecewise exponential with kinks at $z_i=0$ and $z_i=\mu_i$. On each segment,

$$f_i(z_i)=\exp(d+\kappa z_i),$$

with slope κ and shift d determined by the interval and parameters $(\mu_i,\sigma_i,\sigma_{T,i})$. The normalization and required moments can then be expressed in closed form.

For a segment $(z_{\mathrm{lo}},z_{\mathrm{hi}})$:

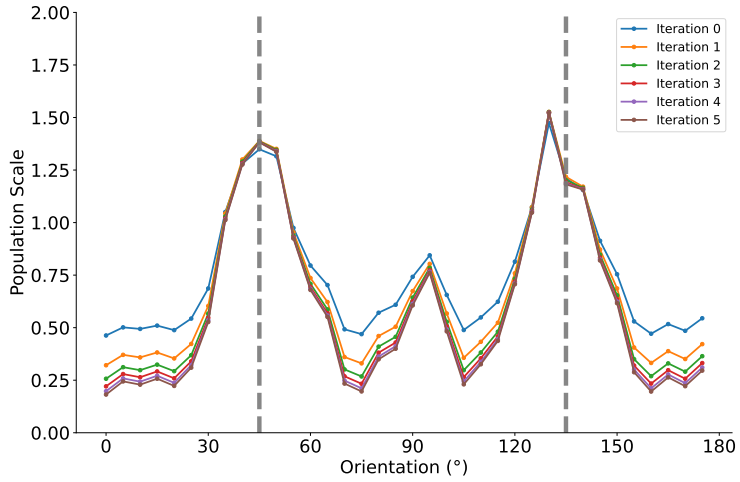

Figure 6: Convergence of the iterative solution of the self-consistency equation for the scale parameters.

- **Normalization:**

$$N_{\text{seg}} = \begin{cases} \frac{e^{d+\kappa z_{\text{hi}}} - e^{d+\kappa z_{\text{lo}}}}{\kappa}, & \kappa \neq 0, \\ e^{d}(z_{\text{hi}} - z_{\text{lo}}), & \kappa = 0 \end{cases}$$

- **First moment:**

$$M_{\text{seg}} = \begin{cases} \frac{z_{\text{hi}} e^{d+\kappa z_{\text{hi}}} - z_{\text{lo}} e^{d+\kappa z_{\text{lo}}}}{\kappa} - \frac{e^{d+\kappa z_{\text{hi}}} - e^{d+\kappa z_{\text{lo}}}}{\kappa^2}, & \kappa \neq 0, \\ \frac{e^{d}}{2}(z_{\text{hi}}^2 - z_{\text{lo}}^2), & \kappa = 0 \end{cases}$$

- **Absolute-value moment:**

$$A_{\text{seg}} = \begin{cases} \frac{|z_{\text{hi}}| e^{d+\kappa z_{\text{hi}}} - |z_{\text{lo}}| e^{d+\kappa z_{\text{lo}}}}{\kappa} - \frac{\text{sgn}(z_{\text{hi}}) e^{d+\kappa z_{\text{hi}}} - \text{sgn}(z_{\text{lo}}) e^{d+\kappa z_{\text{lo}}}}{\kappa^2}, & \kappa \neq 0, \\ \frac{e^{d}}{2}(|z_{\text{hi}}|^2 - |z_{\text{lo}}|^2), & \kappa = 0 \end{cases}$$

By summing across all segments $(-\infty, 0)$, $(0, \mu_i)$, (μ_i, ∞) for $\mu_i > 0$ (and symmetrically otherwise), we obtain

$$N_i = \sum_{\text{segments}} N_{\text{seg}}, \quad \mathbb{E}[z_i] = \frac{\sum_{\text{segments}} M_{\text{seg}}}{N_i}, \quad \mathbb{E}[|z_i|] = \frac{\sum_{\text{segments}} A_{\text{seg}}}{N_i}.$$

To avoid numerical overflow/underflow in practice before evaluating exponentials, we subtract the maximum exponent among all segment endpoints. This does not affect the ratios, but makes the calculation stable.

A.4 Convergence of the Laplace prior fitting

When we performed the iterative determination of the scale components (the solution of 12) we found convergence after five iterations. In Fig. 6, we plot the progression of the scale. The components of the scale vector are categorized into orientations in the same way as the model neurons.

A.5 Alternative prior choices

We investigated alternative hypotheses about the mice's internal prior across D2–D6. As shown in Appx Fig. 7b, when the model uses a 135° NoGo prior, the response profiles differ qualitatively from the measured profiles. A similar mismatch appears when we model an animal that updates

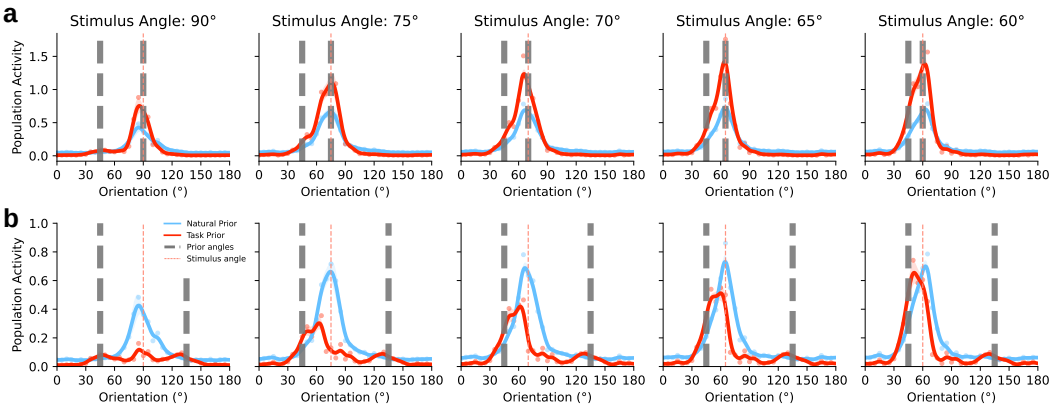

Figure 7: **a**, Model responses when the NoGo prior angle coincide with the test stimulus **b**, D2-D6 model response profiles using constant NoGo prior angles at 135°

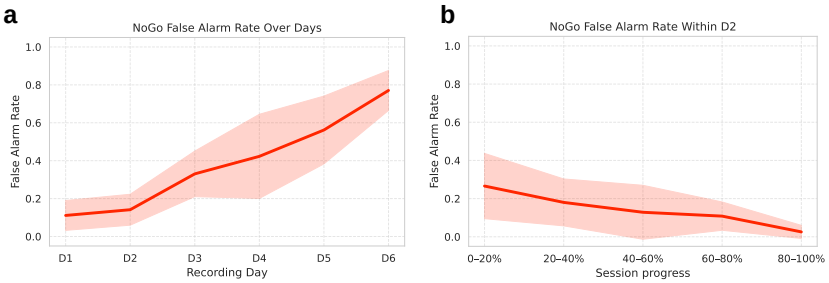

Figure 8: **a**, D2-D6 animal behaviour measured as the false alarm error rate for NoGo stimuli. Across days. **b**, as in *a*, but within D2 over session progression

its prior daily (Appx. Fig. 7a). This is consistent with the false alarm rate (the rate at which the animal licks in response to a NoGo signal) across days: on D2 the animal performs similarly to D1 (Appx. Fig. 8a), indicating adaptation to the new environment, but on subsequent days performance worsens. The animal appears to be improving throughout D2 as it adapts to the new NoGo signal (Appx. Fig. 8b).

A.6 Estimating likelihood and prior in the space of stimulus orientations

Assuming that we observe an orientation posterior in the orientation space when recording a population of neurons, we intend to infer the likelihood and orientation prior in this space from the population response profile. Importantly, this prior–posterior formulation is distinct from that of the VAE framework, which is defined over a high-dimensional neuronal activation space rather than a one-dimensional orientation space.

We focus on D2 responses and focus on one of the modes flanking the dominant mode centered on the stimulus. The flanking mode is a mixture component of the posterior, which is a combination of the contextual prior (pr), and the likelihood (l). Approximating the posterior, likelihood, and prior with a Gaussian in the orientation space, the mean of the mode can be obtained:

$$\mu_{\text{post}} = \frac{\frac{1}{\sigma_{\text{pr}}^2}\mu_{\text{pr}} + \frac{1}{\sigma_{\text{l}}^2}\mu_{\text{l}}}{\frac{1}{\sigma_{\text{pr}}^2} + \frac{1}{\sigma_{\text{l}}^2}} = \frac{\sigma_{\text{pr}}^2\mu_{\text{l}} + \sigma_{\text{l}}^2\mu_{\text{pr}}}{\sigma_{\text{pr}}^2 + \sigma_{\text{l}}^2} \tag{14}$$

In a coordinate system centered around the NoGO stimulus of D2, some of the parameters can be readily determined: $\mu_{\text{l}} = 0$; $\mu_{\text{pr}} = \pi/4$; and μ_{post} can be measured from the population response profile. In a coordinate system centered around the NoGO stimulus of D2, some of the parameters can be readily determined: $\mu_{\text{l}} = 0$; $\mu_{\text{pr}} = \pi/4$; and μ_{post} can be measured from the population

response profile. In a coordinate system centered around the NoGO stimulus of D2, some of the parameters can be readily determined: $\mu_{\mathrm{l}} = 0$; $\mu_{\mathrm{pr}} = \pi/4$; and μ_{post} can be measured from the population response profile.

We rely on the insight that if we know the means of the prior and the likelihood – which we do know – then it is the relative width of the likelihood and the prior that determines the place of the posterior. Let's assume that the prior width is $\sigma_{\mathrm{pr}} = r \cdot \sigma_{\mathrm{l}}$. Then Eq. 14 becomes

$$\mu_{\mathrm{post}} = \frac{r^2\sigma_{\mathrm{l}}^2\mu_{\mathrm{l}} + \sigma_{\mathrm{l}}^2\mu_{\mathrm{pr}}}{\sigma_{\mathrm{l}}^2\left(1 + r^2\right)} = \frac{r^2 \cdot \mu_{\mathrm{l}} + \mu_{\mathrm{pr}}}{1 + r^2} \tag{15}$$

Thus, if we know μ_{post}, then we can calculate r.

We can solve for r in (15) via:

$$\begin{aligned}
\mu_{\mathrm{post}} &= \frac{r^2\mu_{\mathrm{l}} + \mu_{\mathrm{pr}}}{1 + r^2} \\
\mu_{\mathrm{post}}(1 + r^2) &= r^2\mu_{\mathrm{l}} + \mu_{\mathrm{pr}} \\
\mu_{\mathrm{post}} + r^2\mu_{\mathrm{post}} &= r^2\mu_{\mathrm{l}} + \mu_{\mathrm{pr}} \\
r^2\mu_{\mathrm{post}} - r^2\mu_{\mathrm{l}} &= \mu_{\mathrm{pr}} - \mu_{\mathrm{post}} \\
r &= \sqrt{\frac{\mu_{\mathrm{pr}} - \mu_{\mathrm{post}}}{\mu_{\mathrm{post}} - \mu_{\mathrm{l}}}}
\end{aligned}$$

Establishing the width of the likelihood. At this point, we can establish the actual widths of the likelihood and the prior. Consider the bump in the population activity profile when presenting the Go stimulus. There the prior and the likelihood have the same mean. In such a case the posterior is a product of two Gaussians:

$$\mathcal{N}(o; \mu_{\mathrm{Go}}, \sigma_{\mathrm{Go}}) \propto \exp\left(-\frac{1}{2\sigma_{\mathrm{pr}}^2}\left(o - \mu_{\mathrm{Go}}\right)^2\right)\exp\left(-\frac{1}{2\sigma_{\mathrm{l}}^2}\left(o - \mu_{\mathrm{Go}}\right)^2\right) = \tag{16}$$

$$= \exp\left(-\frac{1}{2}\frac{1 + r^2}{r^2\sigma_{\mathrm{l}}^2}\left(o - \mu_{\mathrm{Go}}\right)^2\right) \tag{17}$$

This highlights that the posterior width is

$$\sigma_{\mathrm{post}} = \frac{r}{\sqrt{1 + r^2)}}\sigma_{\mathrm{l}} \tag{18}$$

Using this, we obtain an estimate of the widths of the likelihood and prior from the width of the Go responses.

A.7 Use of LLMs

We occasionally utilized LLMs to refine the paper's text. Additionally, we employed LLM-powered tools for programming, particularly when creating scripts for the figures.