

Remedying uncertainty representations in visual inference through Explaining-Away Variational Autoencoders

Josefina Catoni*, Domonkos Martos*, Ferenc Csikor, Enzo Ferrante, Diego H. Milone, Balázs Meszéná, Gergő Orbán†, and Rodrigo Echeveste†

Abstract—Optimal computations under uncertainty require an adequate probabilistic representation about beliefs. Deep generative models, and specifically Variational Autoencoders (VAEs), have the potential to meet this demand by building latent representations that learn to associate uncertainties with inferences while avoiding their characteristic intractable computations. Yet, we show that it is precisely uncertainty representation that suffers from inconsistencies under an array of relevant computer vision conditions: contrast-dependent computations, image corruption, out-of-distribution detection. Drawing inspiration from classical computer vision, we present a principled extension to the standard VAE by introducing a simple yet powerful inductive bias through a global scaling latent variable, which we call the Explaining-Away VAE (EA-VAE). By applying EA-VAEs to a spectrum of computer vision domains and a variety of datasets, spanning standard NIST datasets to rich medical and natural image sets, we show the EA-VAE restores normative requirements for uncertainty. Furthermore, we provide an analytical underpinning of the contribution of the introduced scaling latent to contrast-related and out-of-distribution related modulations of uncertainty, demonstrating that this mild inductive bias has stark benefits in a broad set of problems. Moreover, we find that EA-VAEs recruit divisive normalization, a motif widespread in biological neural networks, to remedy defective inference. Our results demonstrate that an easily implemented, still powerful update to the VAE architecture can remedy defective inference of uncertainty in probabilistic computations.

Index Terms—Uncertainty, Deep generative models, NeuroAI, Probabilistic Inference, Explaining-Away, Divisive Normalization.

This work was supported by Argentina's National Scientific and Technical Research Council (CONICET) (RE), which covered the salaries of all the authors in Argentina, the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory (GO) and the National Research, Development and Innovation Office under grant agreement ADVANCED 150361 (GO). The authors gratefully acknowledge NVIDIA Corporation for the GPU computing, and the support of Universidad Nacional del Litoral (Grants CAID-PIC-50220140100084LI, 50620190100145LI), ANPCyT (PICT-PRH-2019-00009, 2022-03-00593).

Josefina Catoni, Diego H. Milone and Rodrigo Echeveste are with the Research Institute for Signals, Systems and Computational Intelligence sinc(i) (FICH-UNL/CONICET), 3000 Santa Fe, Argentina (e-mail: jcatoni@sinc.unl.edu.ar; dmilone@sinc.unl.edu.ar; recheveste@sinc.unl.edu.ar).

Domonkos Martos, Ferenc Csikor, Balázs Meszéná and Gergő Orbán are with the Department of Computational Sciences, HUN-REN Wigner Research Centre for Physics, 1121 Budapest, Hungary (e-mail: martos.domonkos@wigner.hun-ren.hu; csikor.ferenc@wigner.hun-ren.hu; meszena.balazs@wigner.hun-ren.hu; orban.gergo@wigner.hun-ren.hu).

Enzo Ferrante is with the Institute of Computer Sciences (UBA/CONICET), 1428 Buenos Aires, Argentina, and also with the Research Institute for Signals, Systems and Computational Intelligence sinc(i) (FICH-UNL/CONICET), 3000 Santa Fe, Argentina (e-mail: eferrante@dc.uba.ar).

* These authors contributed equally.

† These authors jointly supervised this work.

Corresponding author: Josefina Catoni

I. INTRODUCTION

AN adequate representation of uncertainty is key to reliable decision making and sensory integration [1], and thus its importance for artificial systems is increasingly evident as flexibility, adaptation and multimodality become widespread requirements. Maintaining an estimate of the uncertainty associated with inferences permits better reasoning about beliefs [2], generalization [3], and assessment of risks associated with alternative decisions [4]. Furthermore, it allows for more efficient use of acquired knowledge for learning [5]–[7], and downstream computations in a hierarchical processing pipeline, among others. In particular, uncertainty-aware AI systems are argued to be necessary for exploring novel domains [8], [9]. Indeed, probabilistic machine learning promises to deliver tools that are not only capable of inferring best estimates but also the associated confidence in those estimates [10]. In this context, generative models have been advocated as a central framework for probabilistic machine learning [11], [12]. Recently, deep learning has been elevated to accommodate deep generative models, including generative adversarial networks [13], variational autoencoders (VAEs) [14], [15], and diffusion models [16]–[18]. In particular, when learning a latent representation of the data, VAEs provide an explicit representation of the uncertainty of inferences, rendering them a central tool for applications that require probabilistic computations [14], [19]–[21].

A key element of probabilistic computations is that evidence, the likelihood, is combined with expectations, the prior, to interpret observations, yielding the posterior distribution whose width carries information about uncertainty of inferences. Given the central role of uncertainty in VAEs, it is crucial to evaluate to what extent uncertainty representations fulfill elementary requirements set by probability theory. Indeed, the capacity of deep generative models to faithfully represent uncertainty has been questioned [22]. Focusing on visual inference tasks, we first consider contrast, a salient variable of natural image statistics, and identify four fundamental requirements towards principled inference [23]. First, at zero contrast, where the image carries no information, the inferred posterior needs to be equal to the prior (Fig. 1a).

Second, with increasing contrast, observations become progressively more informative and the corresponding posteriors should become increasingly distinct from the prior. Accordingly, the average distance between the means of the posteriors

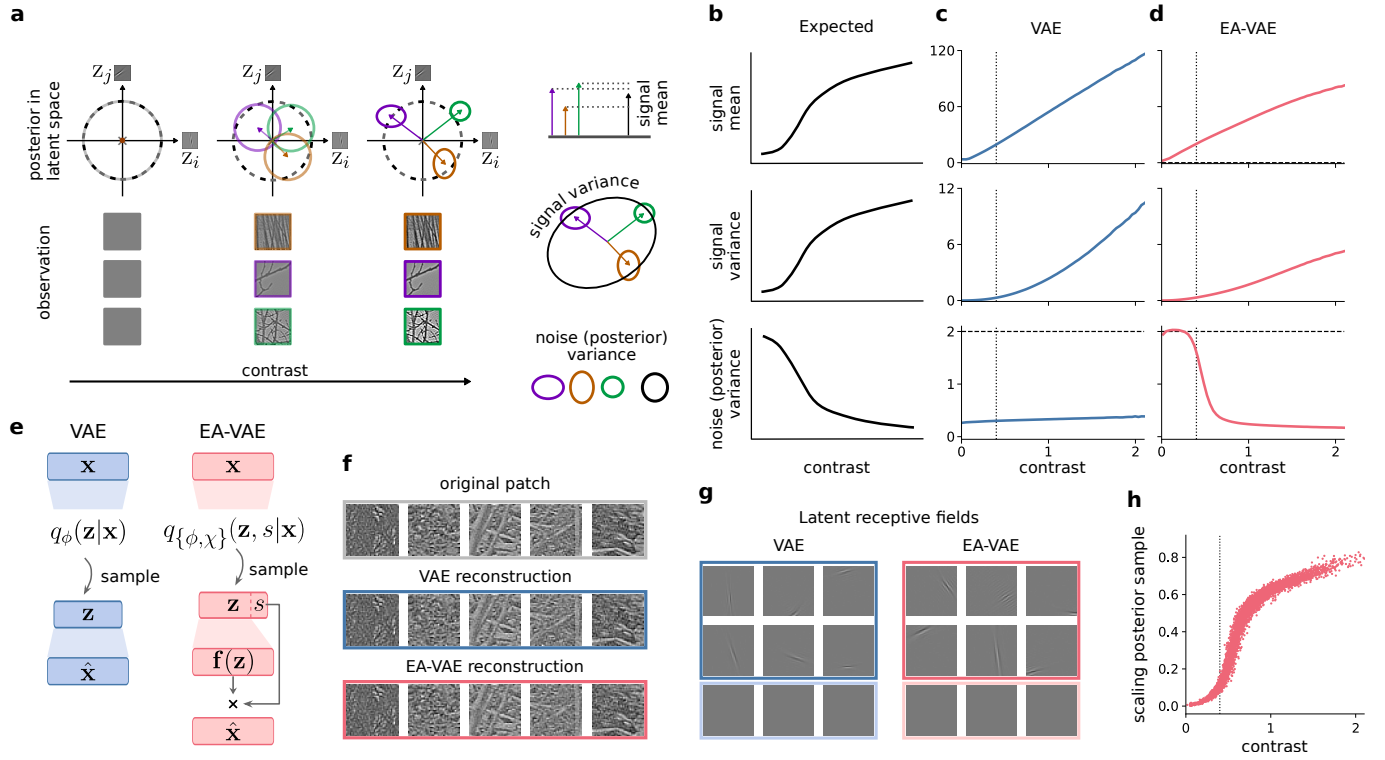


Fig. 1. Characteristic properties of inferred posteriors as contrast is varied. **a**, Expected contrast-dependence of the posterior of a pair of latent variables for three example images (image frame colors matching those of schematic posteriors). As contrast increases, posteriors (colored solid lines) progressively deviate both from the prior (dashed line) and from one another (resulting in *signal variance*). We refer to the average distance between the prior's mean and the posteriors' mean as *signal mean*. Meanwhile, uncertainty estimates, reflected by the width of the posterior, show concomitant changes. We refer to posterior variance for a fixed image as *noise variance* as it does not stem from changes in the signal. **b**, Cartoon of expected contrast-dependence of signal mean, signal variance and noise variance. **c,d**, Contrast-dependence of signal mean, signal variance and noise variance of latent posteriors in natural image-trained VAE (**c**) and EA-VAE (**d**, respectively, Supplementary Section II-B2). Prior statistics and observation noise are indicated in horizontal and vertical dashed lines. **e**, Comparison of the inference and reconstruction process in the VAE and EA-VAE, EA-VAE featuring a multiplicative scaling latent variable, s . **f**, Example image patches and reconstructions. **g**, Receptive filters of example latents in the VAE and EA-VAE. Informative/non-informative latent units are surrounded by high/low saturation colored boxes (Supplementary Section II-B1). **h**, Inferred posterior samples of the scaling variable for individual patches in the EA-VAE model, as a function of the measured contrast of these images.

and the prior, which we term *signal mean*, should increase (Fig. 1a and b, top). Third, more informative observations yield increasing variance in posteriors means (Fig. 1a and b, center), usually called *signal variance*. Finally, with increasing contrast observations become more informative, resulting in decreasing uncertainty about the latent generative component, i.e. decreasing posterior variance, referred to as *noise variance* (Fig. 1a and b, bottom). However, in this work we show that when training a VAE on natural images (Fig. 1f), although contrast-dependence of the posterior mean matches these requirements, contrast-dependent changes in posterior variances are at odds with the promise of a faithful representation of uncertainty (Fig. 1, cf. b and c).

We argue that this signature of inaccurate inference is a consequence of the mismatch between the assumptions about the statistics of latent representations and those characteristic of generative factors underlying naturally occurring inputs. In its basic formulation, latent variables in VAEs are assumed to be either independent or linearly correlated. Natural stimuli, however, have been shown to display strong nonlinear dependencies between latents [24].

We propose a theoretically well-motivated and powerful update to the standard VAE architecture, the Explaining-

Away-VAE (EA-VAE for short). EA-VAE incorporates an inductive bias in the generative model by learning an extra latent variable, which acts as a global multiplicative scaling that is capable of modulating the variance of the data. This extension was inspired by the Gaussian Scale Mixtures (GSM) model [24], a simple yet powerful model of natural images for compression and denoising, which has also been successful in interpreting behavioral and neural data in biological perception [25]–[28]. As recruiting the scaling variable can explain global changes in the input, the introduced scaling variable can replace correlated modulations of latent activations upon changes in image contrast. Hence the term explaining-away in our EA-VAEs.

In this work, we demonstrate that EA-VAEs address a broad set of issues concerning VAEs, beyond producing contrast-dependent posteriors that match the fundamental requirements of probabilistic inference: the reliable representation of the posterior uncertainty contributes to enhanced inference capabilities in a broad set of tasks. First, we show that limited inference capabilities of VAEs are present in canonical computer vision domains such as handwritten character recognition [29] and x-ray medical images [30], which are consistently remedied by EA-VAEs. Second, we demonstrate

that the introduction of the scaling variable in the generative model shapes learned representations such that it promotes disentanglement of latent variables. Third, we demonstrate that EA-VAEs address challenges beyond contrast-dependent inferences, correcting inference for image manipulations that affect uncertainty and addressing a widely-known challenge of VAEs: the detection of out-of-distribution data. Fourth, we show the benefits of a more faithful uncertainty representation through information fusion. Finally, we identify biologically motivated properties of the recognition model of EA-VAEs that complement the inductive bias introduced in the generative model through the scaling latent variable. We demonstrate that improved inference in EA-VAEs is made possible by computations that bear the signatures of a widely known computational motif in biological neural networks: divisive normalization [31], [32]. The results open up multiple novel avenues of research, showing great promise for machine learning applications where uncertainty estimates are of importance, such as information fusion or anticipation of failure modes.

II. METHODS

We start by presenting the formalism of standard VAEs, which we subsequently use to introduce the EA-VAE. VAEs learn about the statistics of data through discovering latent variables, $\mathbf{z} \in \mathbb{R}^D$. Attractively, latent variables can be considered as generative factors underlying data. This relationship between latent variables and data is summarized as a generative model $p(\mathbf{x}|\mathbf{z})$, which is complete with a prior distribution over latent features, $p(\mathbf{z})$. Upon observing a new data point, $\mathbf{x} \in \mathbb{R}^M$, an inference is made by calculating the posterior distribution $p(\mathbf{z}|\mathbf{x})$. Obtaining an exact posterior is in general prohibitive and VAEs establish a highly principled approach to obtain an approximate posterior by learning a recognition model, $q(\mathbf{z}|\mathbf{x})$. VAEs learn in a self-supervised manner, reconstructing the input via the encoder and decoder models, also referred to as the recognition and the generative model. The encoder and the decoder are parametrized in terms of neural networks with parameters ϕ and ψ , respectively. For any input $\mathbf{x}$, the encoder calculates the parameters of the variational approximation of the posterior, $q_\phi(\mathbf{z}|\mathbf{x})$. After taking a sample $\mathbf{z}$ from $q_\phi(\mathbf{z}|\mathbf{x})$, the autoencoder attempts to reconstruct the input by synthesizing $\hat{\mathbf{x}}$ through the decoder, $p_\psi(\mathbf{x}|\mathbf{z})$ (Fig. 1e, left). Depending on the pixel distribution of the data, the likelihood can be a normal or a Bernoulli distribution, with independent pixels in both cases.

The objective function to learn the parameters of the encoder and decoder is formulated as the Evidence Lower Bound (ELBO) [33]. This takes the form

$$\mathcal{L}_{VAE}(\mathbf{x}, \phi, \psi) = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\psi(\mathbf{x}|\mathbf{z})] + \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \quad (1)$$

where the prior, $p(\mathbf{z})$, is usually a standard normal distribution. The first term expresses the reconstruction error, and the second term acts as regularization, penalizing differences between the prior distribution and the approximate posterior. The

hyperparameter β determines the prior regularization weight, and is linked to the assumed pixel observation noise σ_{obs}^2 [34].

The expectation in the first term is not explicitly calculated, but a Monte Carlo approximation is taken with a single sample from the variational posterior. With this approximation, the reconstruction error takes either the form of the cross-entropy or quadratic loss between an input and its reconstruction for the Bernoulli and normal likelihoods, respectively.

A. The explaining-away variational autoencoder (EA-VAE)

In this paper we propose a novel variational inference architecture, specifically designed to improve uncertainty estimates. To that end, the EA-VAE extends the latent space of standard VAEs with a scalar latent variable $s \in \mathbb{R}^+$ that acts multiplicatively in the decoder. That is, the mapping $\mathbf{f}(\mathbf{z})$ in the generative model of a standard VAE is supplemented by s such that the mean of the likelihood in the EA-VAE is obtained as $s\mathbf{f}(\mathbf{z})$ (Fig. 1e, right). The scaling variable is inferred simultaneously with the other latents, $\mathbf{z}$. The multiplicative contribution of s to the likelihood implements a division of labor between latents: while traditional latents $\mathbf{z}$ encode a variety of details about the input, s accounts for overall variance in the pixel space. Intuitively, the scaling variable can represent global modulations in an image without changes in the content of the image. This division of labor has the consequence that s is directly in charge of modulating the model's uncertainty about the input. Large values of s encourage the network to represent the input with high fidelity and certainty, while small values of s encourage the posterior of $\mathbf{z}$ to collapse to the prior, reporting maximal uncertainty (see mathematical derivation in Supplementary Section I).

For EA-VAEs, the variational posterior is a joint distribution: $q_{\{\phi, \chi\}}(\mathbf{z}, s|\mathbf{x})$, where ϕ and χ denote the parameters of the components of the encoder. A widespread approach for VAEs is to learn a variational posterior that assumes independence among the latents. Accordingly, the joint variational posterior of EA-VAE was broken down into two components, $q_\phi(\mathbf{z}|\mathbf{x})$ and $q_\chi(s|\mathbf{x})$. As s is meant to account for the joint modulation of pixel intensities, a positive value is justified. As shown later, the results are robust against the specific choice of the variational posterior of s . We note that whenever a VAE was contrasted with an EA-VAE, the total number of latent variables was preserved to ensure a fair comparison.

The extended generative model yields an updated training objective, that is, an updated ELBO:

$$\mathcal{L}_{EA-VAE}(\mathbf{x}, \phi, \psi, \chi) = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} [\log p_\psi(\mathbf{x}|\mathbf{z}, s)] + \beta_1 D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + \beta_2 D_{KL}(q_\chi(s|\mathbf{x})||p(s)). \quad (2)$$

Here β_1 and β_2 denote the relative strengths of the prior regularization terms of $\mathbf{z}$ and s and respectively. For reconstruction, both $\mathbf{z}$ and s are sampled from their respective posteriors before decoding.

Next, we present the specific VAE and EA-VAE architectures investigated in the paper.

B. VAEs for natural images

We explored the contribution of the scaling variable to inference in VAEs using an architecture that featured a linear generative model. In this model the scaling variable, s , can be studied in detail as it has a well-defined role for the representation of contrast. Our analysis is aligned with earlier accounts by adopting a large-dimensional latent space, fully connected layers, and a Laplace prior [35], [36]. We trained the model on gray-scale natural image patches.

The Laplace prior serves a sparse coding approach, which has historically been taken to capture natural image statistics and recover the receptive field properties of neurons in the primary visual cortex [37]. We hence employ these models as a starting point for our analysis (though we will thoroughly analyze VAEs with Gaussian priors in later sections). To ensure that the KL divergence in the ELBO is analytically tractable, we employed the Laplace distribution for the variational posterior too: $q_\phi(\mathbf{z}|\mathbf{x}) = \text{Laplace}(\mathbf{z}; \boldsymbol{\mu}(\mathbf{x}), \mathbf{b}(\mathbf{x}))$, with $\boldsymbol{\mu}(\mathbf{x})$ and $\mathbf{b}(\mathbf{x})$ being the respective mean and scale parameters. The generative model for a given $\mathbf{x}$ is parameterized by a multivariate uncorrelated Normal distribution $p_\psi(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}; \mathbf{f}(\mathbf{z}), \boldsymbol{\sigma}_{\text{obs}}^2)$, such that $\mathbf{f}(\mathbf{z}) = \mathbf{A} \cdot \mathbf{z}$ is a linear transformation, and σ_{obs} is the observation noise (see Supplementary Section II-A for detailed cost function)

The EA-VAE extension preserves the previous model structure and incorporates the scalar multiplicative latent variable s (cf. Fig. S1 a and b). To avoid degeneracy in the image reconstruction we assume s to be positive. We implemented a range of variants for this: a softplus activated Laplace distribution, a Log-normal distribution, and a Gamma distribution (see Supplementary Section II-A for detailed cost functions). We used the first variant as a default and used the variants as controls.

C. VAEs for other classic computer vision domains

We studied the properties of EA-VAEs on standard computer vision datasets (MNIST, ChestMNIST, FashionMNIST, NIH-ChestXray14, Imagenet, see details in Section III-A). The corresponding variational autoencoders were not restricted to linear decoders and, as is common practice, convolutional layers were also considered. As is standard in the field, the variational posterior was a multivariate uncorrelated Normal distribution $q_\phi(\mathbf{z}|\mathbf{x}) = \text{Normal}(\mathbf{z}; \boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\sigma}^2(\mathbf{x}))$, with $\boldsymbol{\mu}(\mathbf{x})$ and $\boldsymbol{\sigma}^2(\mathbf{x})$ the respective mean and variance vectors. In the case of the EA-VAE the variational posterior of the multiplicative variable, s , was log-normal: $q_\chi(s|\mathbf{x}) = \text{Log-normal}(s; \nu(\mathbf{x}), \xi^2(\mathbf{x}))$. Priors $p(\mathbf{z})$ and $p(s)$ come from the same family as their respective variational posteriors but with zero mean and diagonal unit variance.

When studying the models on the MNIST dataset, we assumed a Bernoulli distribution in the decoder. This choice is justified by the largely black-and-white distribution of pixel intensities of digits. For other datasets (see details in Section III-A), the decoder was parameterized by a multivariate uncorrelated Normal distribution (Supplementary Section II-A for detailed cost functions).

III. EXPERIMENTAL SETUP

A. Datasets

1) *Natural Images*: Following [38], we used the Van Hateren dataset, composed of natural image patches of 40×40 pixels each [39] (6.4×10^5 images for training and validation and 6.4×10^4 for test). The first set was further divided into 80% for training and 20% for validation purposes. Image patches had already been preprocessed, whereby $100(1 - \frac{\pi}{4})\%$ of the higher frequency PCA components had been discarded. By means of this filtering, the effective dimensionality of the image space had been reduced to 1,256. Note that the patch size remains unaltered. Additionally, for the case with a Gamma scaling variable, the distribution of pixel intensities was rescaled (dividing by an α constant) such that its mean ± 3 standard deviations fit the range $[0, 1]$, to have the same range for all domains. This was important for the post-training evaluations. Finally, the mean pixel intensity of each patch was subtracted before the image enters the encoder network.

2) *NIST datasets*: We employed three popular NIST datasets: MNIST [29], ChestMNIST [30] and FashionMNIST [40], which contain standardized grayscale images for digits, chest radiographs and fashion items, respectively. Each original 28×28 pixelsize image was resized to 32×32 due to the specific used architecture (Section III-B). The pixel intensities for these canonical datasets are in the range $[0, 1]$.

3) *Contrast augmented NIST datasets*: Augmented versions of classic MNIST, ChestMNIST and FashionMNIST datasets were synthetically generated in order to have a rich variety in image contrast. We refer to these datasets as cMNIST, cChestMNIST and cFashionMNIST, respectively. Each image in the original dataset was first subtracted by 0.5, so the pixel intensities move to the range $[-0.5, 0.5]$. Then multiplied by a constant c sampled from a Log-normal distribution, and then intensity was corrupted with a white noise vector $\boldsymbol{\eta}$, sampled from a Gaussian distribution of null mean and σ_{obs} standard deviation

$$\mathbf{x}^{\text{aug}} = c \left(\mathbf{x}^{\text{ori}} - \frac{1}{2} \right) + \boldsymbol{\eta}. \quad (3)$$

This process was repeated such that cMNIST, cChestMNIST and cFashionMNIST datasets resulting in a 10-fold increase in the size of the dataset. σ_{obs} values for each dataset are detailed in Table S2.

4) *Shuffled pixels datasets*: For testing purposes in the out of distribution experiments, all testing subsets (Van Hateren, MNIST, ChestMNIST, cMNIST, cChestMNIST, cFashionMNIST) were used to generate synthetically pixel shuffled datasets. Given a specific testset of size $(N_{\text{img}}, N_{\text{pix}})$, the pixel-shuffled dataset is generated by performing random permutation of pixels such that pixels will get permuted only with pixels that are in the same location. By doing this, the pixel-shuffled dataset would have the same likelihood of observation, yet the information present in the shuffled image will differ from the original datasets, constituting a useful baseline for out-of-distribution detection.

5) *Higher-resolution datasets*: To evaluate uncertainty representations in more complex settings we resorted to two further datasets. Firstly, Imagenet [41], [42], a widely used benchmark for image classification, from which we selected all three categories of human images from the standard validation set: ‘Ball Player’, ‘Groom’ and ‘Scuba Diver’, as well as three object categories: ‘Desk’, ‘Wall Clock’ and ‘Bookcase’. Secondly, we employed the NIH-ChestXray14 dataset [30]. This is a large-scale public dataset of frontal chest radiographs released by the U.S. National Institutes of Health. It is composed of 112,120 images from 30,805 unique patients. The official dataset was split into 70% to train, 10% for validation and 20% to test. Images were resized from 1024×1024 into 224×224 , which is standard processing in medical image analysis. Same as NIST datasets, the pixel intensities for these images were set to the range $[0, 1]$.

B. Implementation and training details

All results presented in the main text correspond to models that were trained in Python with Pytorch 2.x. All models were trained through Pytorch’s Adam optimizer, with learning rate (lr) and weight decay (wd) as detailed in Supplementary Table S1 and Table S2, and all other Adam parameters set to Pytorch’s default. We applied early stopping so that the parameters were finally set at the epoch with best performance on the validation images.

Following [38], beta-annealing was considered to train models on natural image patches: we set $\beta_1 = 0.01$ for the first 100 epochs, linearly increasing to the final $\beta_1 = 1$ over the next 100 epochs. We observed that in the case of the EA-VAE beta annealing had little effect in the training outcome. The final value $\beta_1 = 1$ was determined exploratorily such that the decoder’s reconstructions did not present visibly strong deviations from the original input. Results are robust with respect to changes in these hyperparameters, with no abrupt qualitative changes.

To rule out the possibility of our findings being contingent on specific architectural choices, we explored three variants of architecture and scaling variable family for the EA-VAE, describing below the central model presented in the main text in which the scaling variable is modeled as a softplus activated Laplace distribution (See control variants details in Supplementary Section II-A). The model implements the scaling variable s with minimal update to the standard VAE architecture. Here s is coming from the same family of probability distributions of the regular VAE latents, $\mathbf{z}$ but transformed by a monotonous, invertible function to the positive half-space, by use of a Softplus function. The homogeneous formulation of different latent variables motivates a joint encoder for $\mathbf{z}$ and s , yielding a joint variational posterior $q_{\phi, \chi}(\mathbf{z}, s | \mathbf{x})$. We refer to this version of the model as the *homogeneous* version (see Fig. S1a for detailed architecture). To avoid collapse in the initial phase, a stronger regularization was applied on the scaling variable in these models for the course of the first 100 epochs ($\beta_2 = 10$), which linearly decreased to $\beta_2 = 1$ during the next 100 epochs.

To train the VAE and EA-VAE on the NIST and NIH-ChestXray14 datasets, we also incorporated convolutional lay-

ers to parametrize the encoder and decoder components, since they are commonly employed in these domains (see Fig. S1c and Fig. S1d for detailed architectures). In this case we used latent dimension D much lower to input size, as a classic dimensionality reduction model (See Supplementary Table S2 for complete hyperparameter values).

For the contrast augmented cMNIST, cChestMNIST and cFashionMNIST models, the architectures were the same as NIST datasets, except for the last layer which lacks the sigmoid activation, corresponding to mean square reconstruction error (see Fig. S1c for detailed architecture). β_1 and β_2 were fixed by Eq. S26 given the observation noise in the augmentation process (See Supplementary Table S2).

All models were trained either on a GPU NVIDIA TITAN V with a 64GB RAM memory (PC: 1) or on a NVIDIA GeForce RTX 2080 Ti with a 64GB RAM memory (PC: 2). As a reference, the time taken to achieve each respective best epoch on PC 1 in this setting was of 62.2 hours for the VAE vs 67 hours for the EA-VAE trained on Van Hateren natural image patches. That is, the improvement in uncertainty representation in the EA-VAE comes at an expense of roughly an additional 10% computational time. In the case of the models trained on MNIST and ChestMNIST employing convolutional layers, all models took around one hour to complete the 500 epochs.

C. Post-training evaluation of VAEs and EA-VAEs

As a general form to evaluate the models’ representation of uncertainty, we computed neurons responses in the latent space for different observations. For all experiments involving uncertainty in the latent space, we define the uncertainty of the network for a given input $\mathbf{x}$ as the standard deviations of the posterior $q_{\phi}(\mathbf{z} | \mathbf{x})$ averaged across latent dimensions,

$$u(\mathbf{x}) = \text{noise std.} = \text{mean}_j \sigma_j(\mathbf{x}). \quad (4)$$

We call this as well the *posterior width*.

When evaluating models for inference on natural images, latent units were discerned as informative or non-informative according to their contribution to the dataset reconstruction. All experiments were computed using only the informative latent dimensions (experimental details in Supplementary Section II-B).

To assess uncertainty representations in classic computer vision domains, models were evaluated in a series of experiments that include out-of-distribution detection, image morphing and image corruption (experimental details in Supplementary Section II-C).

IV. RESULTS

A. Inference on natural images

To obtain a first intuition for the role of the explaining-away variable s , we analyzed natural image patches, where contrast is inherently present in the data, and therefore modulates uncertainty of inferences about image content (Fig. 1f, top row and Section III-A). First, we trained both a standard VAE and an EA-VAE under identical conditions: employing the same family of distributions, data and hyper-parameters. The only

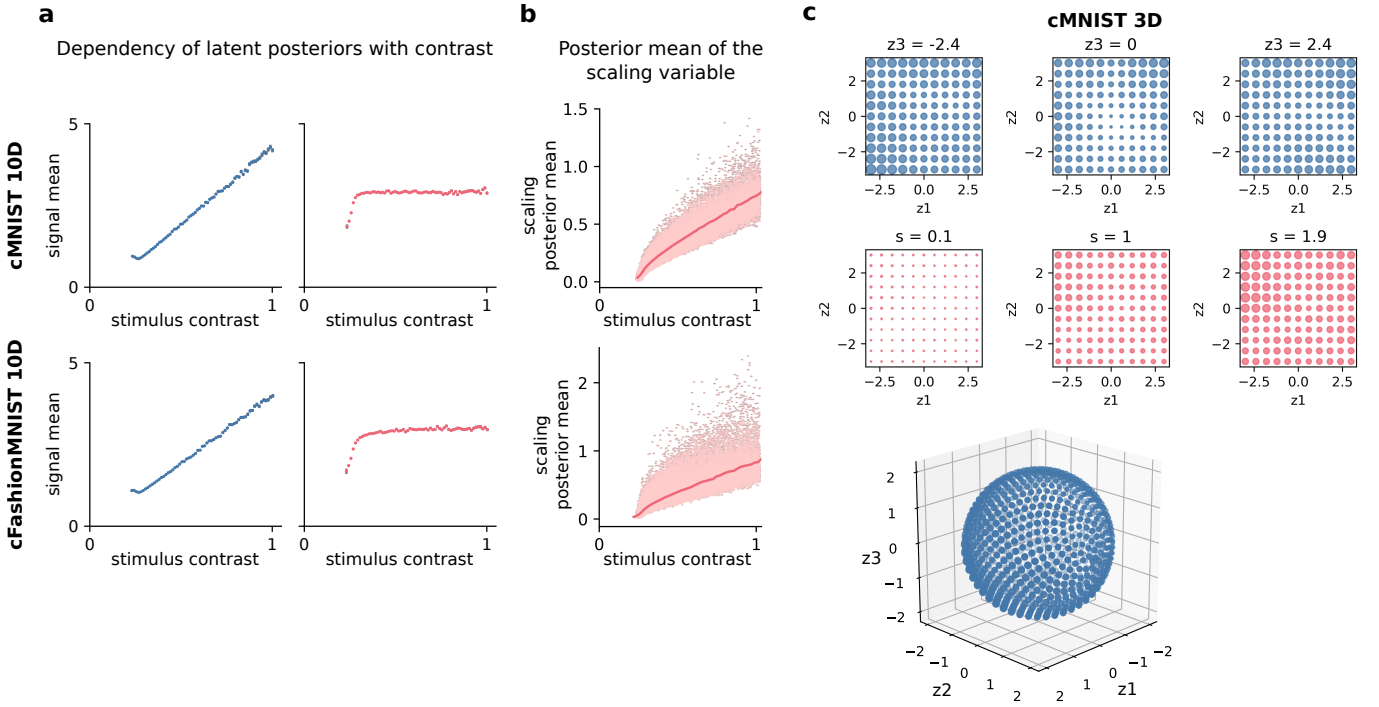


Fig. 2. **Characterization of learned representations for different training data sets.** Beyond natural image patches, contrast-augmented versions of the MNIST and Fashion MNIST databases are shown. **a**, Latent posterior signal mean as a function of the measured contrast of a large set of test images for both the VAE and the EA-VAE models. **b**, Inferred posterior mean of the scaling variable for individual images (dots) in the EA-VAE model, as a function of the measured contrast of these images. Average mean shown as solid line. **c**, Direct comparison of the learned representation of a cMNIST-trained standard VAE (blue) and an EA-VAE (red) both featuring a three-dimensional latent space. *Top panels*: Cross-sections of the latent space (see labels above panels). *Dots* show the contrast of images generated from individual states of the latent space, with dot size proportional to the contrast of the generated image. Note that cross sections for the EA-VAE correspond to different levels of the scaling variable. *Bottom panel*: Contrast of images generated from the three-dimensional latent space of the standard VAE model at a fixed distance from the origin. The fixed distance was 2SD of the prior distribution.

difference is that in the EA-VAE we select one of the latent variables to play the role of a scaling variable, multiplicatively affecting the output. The total number of latent variables is hence the same and the model complexity does not increase in EA-VAE.

1) *Latent representations*: We find that both the VAE and the EA-VAE can be successfully trained on this task in terms of their capacity to reconstruct natural images (Fig. 1f and Fig. S2a). The models also exhibited qualitatively similar latent representations which were consistent with sparse-coding of natural images, with a fixed set of latents showing localized and oriented filters, and a smaller fraction showing unstructured filter properties (Fig. 1g and Fig. S2b).

2) *Latent uncertainty*: Characterization of latents through filter properties informs us how the *mean* of the posterior is modulated by image properties. In order to characterize the relationship between how stimulus content and contrast selectively contribute to latent activations, we characterize the posteriors through the signal mean, signal variance, and noise (or posterior) variance using natural image patches that display changes in their effective contrast level. These quantities are estimated in all cases as averages over the z latents and over images binned according to true contrast (Supplementary Section II-B2). Both VAEs and EA-VAEs showed increased signal mean and signal variance with higher contrast, as

posteriors moved away from the prior and became more distinct (Fig. 1, cf. c and d, top and middle panels). However, in standard VAEs, posterior uncertainty (noise variance) does not decrease with increasing contrast (Fig. 1c, bottom). Particularly, as contrast tends to zero, and local orientations become progressively harder to distinguish, posterior uncertainty fails to converge to the prior uncertainty. In contrast, EA-VAEs display a decreasing uncertainty for higher contrasts. Consistent with the assumptions of the generative model, as contrast decreases below the assumed pixel noise level, the reported noise variance in the EA-VAE saturates at the level of prior uncertainty (Fig. 1d, bottom panel). This observation is also consistent with the fact that the latent s is inferred as close to zero below this level (Fig. 1h). Moreover, as expected given its multiplicative role, s grows monotonically with image contrast in EA-VAEs trained on natural images (Fig. 1h). These results suggest that the explaining away variable s could be playing the role of a signal-to-noise ratio estimate, with low values corresponding to an interpretation that the image is mostly noise, and consequently the posteriors for other latents relax towards the prior (see Supplementary Section I-A). We note that the results obtained for this homogeneous variant of the EA-VAE were qualitatively analogous to the ones obtained for the other EA-VAE variations: with the scaling variable parametrized by lognormal or gamma distributions (Fig. S3),

and when changing β_1 and β_2 hyperparameters (Fig. S13).

We note that the reconstruction term of the ELBO in Eq. 1 is calculated using a single sample from the variational posterior. One may suspect this could lead to underestimation of posterior variance, hence contributing to failed inference in VAEs. We therefore investigated whether uncertainty estimates could be corrected by a VAE variant in which multiple samples are taken to evaluate the ELBO. For this, we implemented the Importance Weighted Variational Autoencoder (IWAE) [43]. In particular, we wanted to assess whether IWAEs could improve uncertainty representations in VAEs without the need to modify the generative model. We trained the IWAE on the same dataset and conditions and found no evidence of multi-sample ELBO estimates correcting the contrast-dependence of the variational posterior (Fig. S11).

B. Improved inference of EA-VAE

Having established a first intuition for the role of an explaining away variable in VAEs, we next explored whether the improved uncertainty representation in EA-VAEs may extend to other classical domains of computer vision. For this, we used standard datasets of computer vision: the MNIST, ChestMNIST, and FashionMNIST datasets, which we adapt for our investigation. To this end, we elevate these data sets with contrast-modulation (Section III-A, referred to as cMNIST, cChestMNIST and cFashionMNIST, respectively).

We trained VAEs and EA-VAEs on cMNIST and cFashionMNIST datasets. In both domains we found that the signal mean grows unbounded with contrast in VAEs, while it quickly saturates for EA-VAEs (Fig. 2a). Conversely, the explaining away variable was strongly modulated by contrast in the EA-VAE for both of these extended datasets (Fig. 2b). This suggests that s contributes to the disentanglement of stimulus content from contrast in EA-VAE.

Next, we sought to understand how EA-VAE's inductive bias shapes its learned representation. For this, we trained three-dimensional latent space versions of VAE and EA-VAE models, whose latent space can be directly visualized. Images generated from a grid in the latent space of the VAE show strong modulations of image contrast in all latents (Fig. 2c, top), with equal-contrast images residing approximately on a sphere (Fig. 2c, bottom). In contrast, images generated from the regular latent variables of the EA-VAE (z_1 and z_2) showed homogeneous contrast at any given value of the explaining away latent, s (Fig. 2c, middle). In summary, the explaining away variable contributes to the disentanglement of contrast from other latent factors in EA-VAEs, which does not naturally occur in standard VAEs (though see [44]).

C. Out-of-distribution inferences with EA-VAE

After establishing that the explaining-away variable implements a direct inductive bias for learning about contrast, we next explored the contribution of this inductive bias to other computer vision tasks. In particular, VAEs were shown to struggle with detecting out-of-distribution (OOD) data, namely data points that come from a distribution whose statistics do not match the data distribution the VAE was trained on

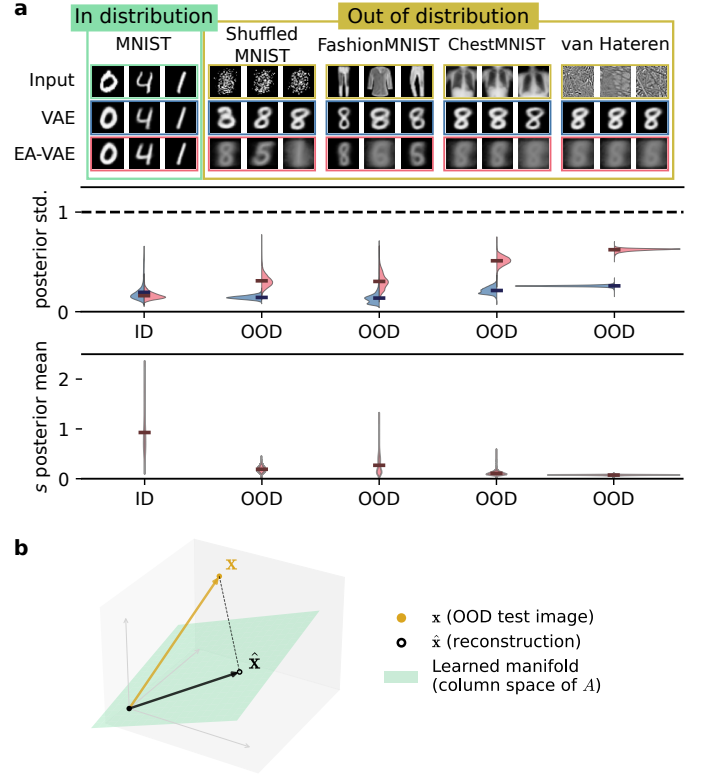


Fig. 3. **Characterization of posterior width for out-of-distribution experiments in models trained on MNIST.** a, Posterior width and the magnitude of the scaling variable for in-distribution (ID) and out-of-distribution (OOD) examples both for standard VAE (blue) and EA-VAE (red). Top: Examples of images from the same domain (light green frames) and different domains (mustard frames) as inputs, and reconstructions by the standard VAE (second row) and EA-VAE (third row) respectively. Bottom: uncertainty quantified by the posterior width and mean of scaling variable posterior for individual ID or OOD images. As a reference, a dashed black line indicates the prior uncertainty. b, Learned linear manifold spanned by ID training data and the projection $\hat{x}$ of an OOD test data point x to the learned manifold.

[22]. Indeed, testing a VAE trained on the standard MNIST dataset shows a striking pattern: reconstructed images of a standard VAE when presenting shuffled pixels, or images coming from other datasets are surprisingly similar to in-distribution reconstructions (Fig. 3a). Accordingly, posterior variances show no effect of increased uncertainty (Fig. 3a). Equivalent results were found when training on the other NIST datasets (see Supplementary Section II-C1 and Fig. S5).

To better understand how the introduction of an explaining-away variable might address this issue, we sought to find an analytical treatment of OOD inferences in EA-VAEs. As the scaling variable s is directly related to posterior uncertainty (Supplementary Section I-A), we focus on the behavior of this variable. Assuming a linear generative model where $\mathbf{A}$ maps latents ($\mathbf{z}$) to observed variables $\mathbf{x}$, it can be shown (Supplementary Section I-B) that the explaining-away variable in an EA-VAE can be expressed as a function of the stimulus contrast $\|\mathbf{x}\|^2$:

$$s^2(D + \log s) \propto \|\mathbf{x}\|^2 \cos(\mathbf{x}, \hat{\mathbf{x}})^2, \quad (5)$$

where the proportionality constant is determined by $\mathbf{A}$, D is the dimensionality of the latent space, and the reconstruction $\hat{\mathbf{x}}$

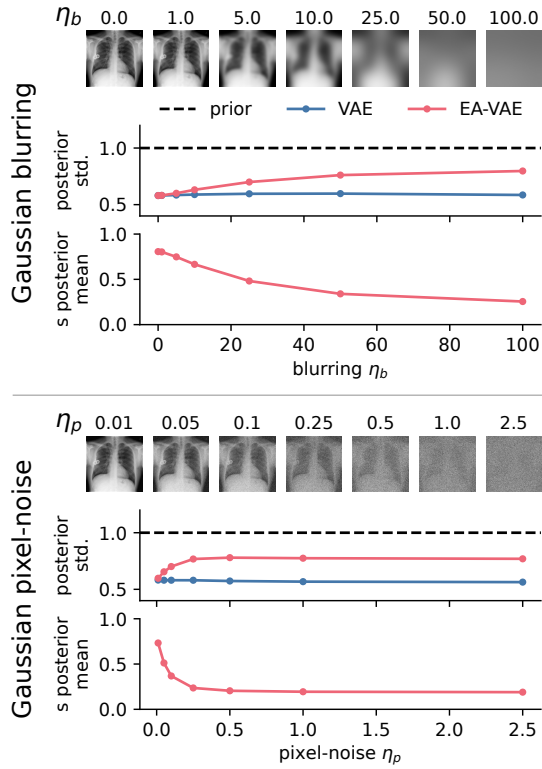


Fig. 4. **Uncertainty representations in high-dimensional medical images.** Posterior width and magnitude of the scaling variable of standard VAE and EA-VAE models trained on NIH-ChestXray14, for images increasingly corrupted either by Gaussian blurring (top) or additive pixel-noise (bottom).

corresponds to the projection of $\mathbf{x}$ to the learned linear manifold spanned by in-distribution data (Fig. 3b). Importantly, the second factor on the right hand side of Eq. 5 is the cosine similarity between the input data points and the corresponding reconstructed data points. This provides an intuitive insight: when presenting an image that is an outlier to the learned manifold spanned by in-distribution data, the cosine similarity will be low and hence so will be s . In turn, lower s values encourage posteriors to collapse to the prior for those inputs, increasing posterior variances (Supplementary Section I-A). In summary, the EA-VAE is expected to display higher variance for OOD images. An EA-VAE that features a linear decoder confirms these analytical results (Fig. S4), and the general setting that features a nonlinear decoder shows that, in contrast with the standard VAE, the uncertainty of inferences in the EA-VAE is systematically modulated by mismatched training and test statistics (Fig. 3a).

Inspired by this insight, and encouraged by the results obtained on simple MNIST datasets, we tested whether the ability of EA-VAEs to detect OOD examples could extend to a more complex setting such as distinguishing human versus non-human images from Imagenet [41], [42] (Supplementary Section II-D). Concretely, we trained VAEs and EA-VAEs on images from one human category and evaluated them on images either from another human category or on images from an object category. Interestingly EA-VAEs showed more dissimilar distributions of uncertainty between human and object images than VAEs, hinting at a more structured semantic

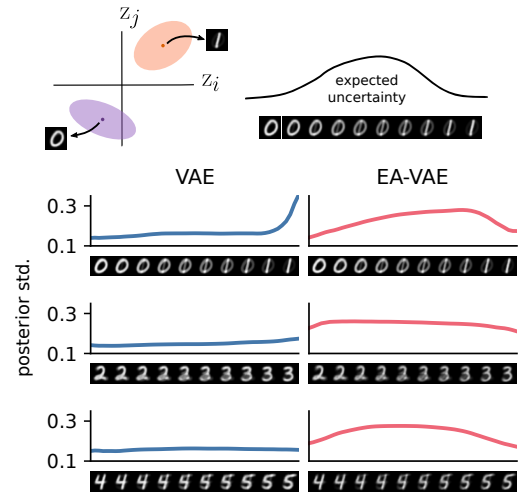


Fig. 5. **Characterization of posterior width for morphing experiment that affect inference uncertainty for models trained on MNIST.** Top left: Average posterior for images with the same label ('0' in purple and '1' in orange). Top right: Gradual morphing between representative image samples from digit '0' and digit '1'. Illustration of the expected changes of the posterior width is shown above. Bottom: Posterior widths for the morphed digits with standard VAE (left) and EA-VAE (right).

representation of uncertainty (Fig. S10). These results hint at a promising avenue of research into OOD detection with EA-VAEs in broader more complex settings.

D. Image corruption and uncertainty

OOD experiments confirmed that the scaling variable introduced in EA-VAE enables it to reflect the mismatch between the training data and the input through posterior uncertainty. Another form of relevant mismatch to be flagged by models concerns images of lower quality or subjected to corruption that may impede correct inferences. To assess this, we analyzed the representation of uncertainty in VAEs and EA-VAEs trained on Chest x-ray images. In order to showcase the scalability of EA-VAEs to larger scale problems we present here the results corresponding to higher resolution images from the NIH-ChestXray14 (Section III-A), though equivalent results were found for the lower definition images from ChestMNIST employed in the OOD analysis (Fig. S7).

We progressively corrupted images, rendering them increasingly uninformative, in two ways: by applying a Gaussian blur (Fig. 4 top) and by adding Gaussian pixel noise (Fig. 4 bottom) (Supplementary Section II-C2). In both cases the EA-VAE displays increasing uncertainty as the image quality deteriorates. This starkly contrasts with the behavior of standard VAEs, where latent uncertainty remains almost constant at low values even as the original image is no longer recognizable from the corrupted one (cf. blue and salmon lines, Fig. 4). This behavior was consistent with the observation that, when faced with an intentionally uninformative image obtained by averaging images over the entire dataset, EA-VAEs report high uncertainty, while standard VAEs do not (Fig. S8).

E. Morphing and uncertainty

When dealing with a domain containing distinct stimulus categories, the latent space is expected to adopt a non-

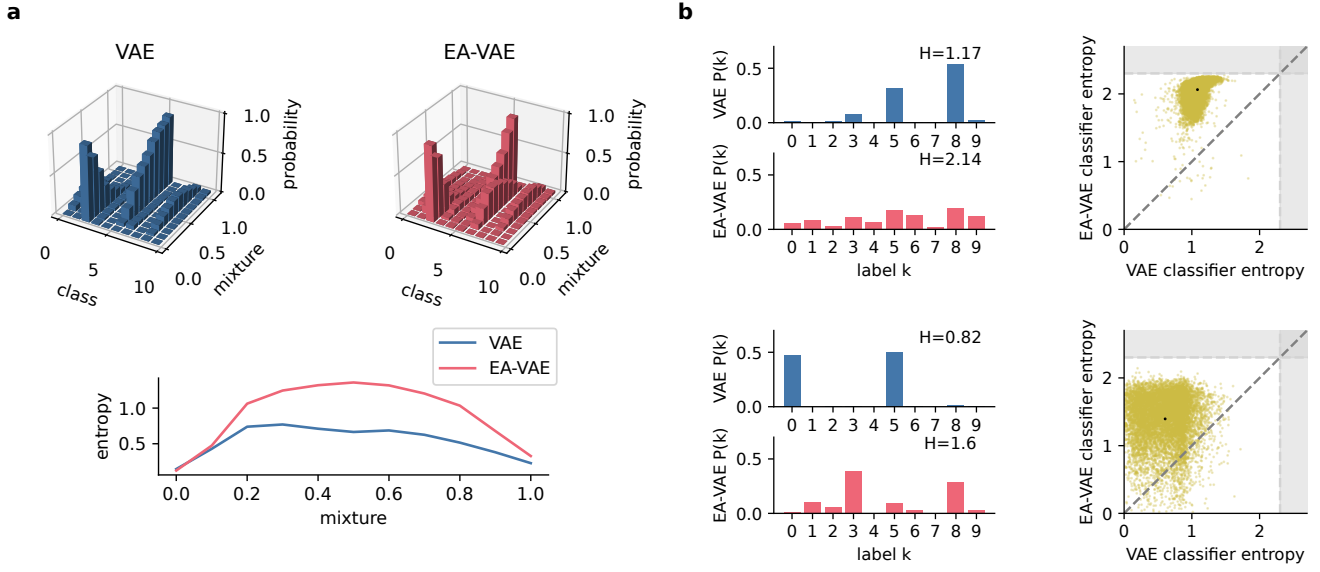


Fig. 6. **Predictive uncertainty in a downstream classification task.** Characterization of an MLP trained for classification task of MNIST handwritten digits through latent posterior representations of a VAE and an EA-VAE. **a**, Behavior of the trained MLP when gradually morphing between image samples from digit '2' to digit '5'. *Top panels*: Histograms of output predicted probabilities averaged for different samples of the digits when trained with each model latent representations. *Bottom panel*: Entropy of predicted probability distribution as a function of combination weight of labels. **b**, Behavior of trained MLP when testing with OOD ChestMNIST (*Top*) and pixel-shuffled MNIST (*Bottom*) images. *Left columns*: Histograms show the distribution of predicted probabilities of labels of an example image, averaged for different samples of the posterior distribution of that image, and its corresponding entropy. *Right column*: Entropy of predicted probability distribution for the MLP trained with each model representations. Each yellow dot represents a single x-ray image, with mean value in black. Identity in dark gray slashed line. Shading corresponds to the upper limit determined by the uniform distribution.

overlapping representation in which inferences about individual images of a given category yield a low uncertainty estimate after learning. In contrast, inference is expected to reflect an increased uncertainty for images that might belong to multiple categories. Note that as VAEs are trained in an unsupervised fashion, these do not rely on category labels during training. Still, we argue that distinct categories correspond to learned subdomains of the image space and morphing systematically drives images to domains of the pixel space with more limited exposure during training. Systematic changes in uncertainty due to ambiguous image categories can be investigated in the case of MNIST-trained models by assessing posterior uncertainty for images of simple digits and for images that are produced by merging two digits from different classes. We conducted this numerical experiment for each of the 45 pairs of different digits and measured the uncertainty along different levels of interpolations between the digits (Fig. 5, and Fig. S6, for alternative pairings). We found a consistent modulation of uncertainty for the EA-VAE model, featuring a clear maximum at an intermediate interpolation level in all except one of the digit pairs (see also Supplementary Section II-C4). In contrast, no consistent peak was identified for the standard VAE model: only 31% of all cases featured a central maximum (14 out of 45 in total).

Methods such as Monte-Carlo (MC) Dropout [5], have often been proposed to achieve better uncertainty estimates in deep learning, by keeping drop-out layers active at inference time and incorporating the additional variance introduced by drop-out. As a control experiment, we studied whether MC Dropout would suffice to remedy uncertainty representations without affecting the generative models. We repeated both

morphing and out-of-distribution experiments and found this did not correct uncertainty estimates in these cases (Fig. S12). Controls such as this and that of IWAEs provide further support for the idea of a fundamental mismatch in the basic assumptions VAEs make about the statistics of the data, which require addressing how the generative model itself is modeled.

F. From latent to predictive uncertainty

Next, we investigated whether the benefits from improved latent uncertainty representations in EA-VAEs also transferred to improved predictive uncertainty in a downstream classification task. Optimal information fusion relies on computing with proper uncertainty estimates, and VAEs are often employed for dimensionality reduction as a previous step to classification. Hence, when multiple sources need to be combined, as it is often the case in medical settings, improved uncertainty estimates could be beneficial.

To investigate this, multilayer perceptrons (MLPs) were trained to classify MNIST digits using samples from the VAE (or EA-VAE) posterior latent representations (Supplementary Section II-C6). Two experiments evaluated the impact of these representations on information fusion and out-of-distribution examples. First, using a set of morphed digits, we compared the predictive uncertainty of MLPs, quantified as the entropy of the output distributions. The EA-VAE demonstrated significantly higher uncertainty (entropy) for intermediate images compared to the VAE ($p_{\text{value}} < 10^{-15}$). (Fig. 6a and Fig. S9). Second, testing on out-of-distribution inputs (ChestMNIST and pixel-shuffled MNIST), the EA-VAE again yielded higher entropy predictions than the VAE (Fig. 6b).

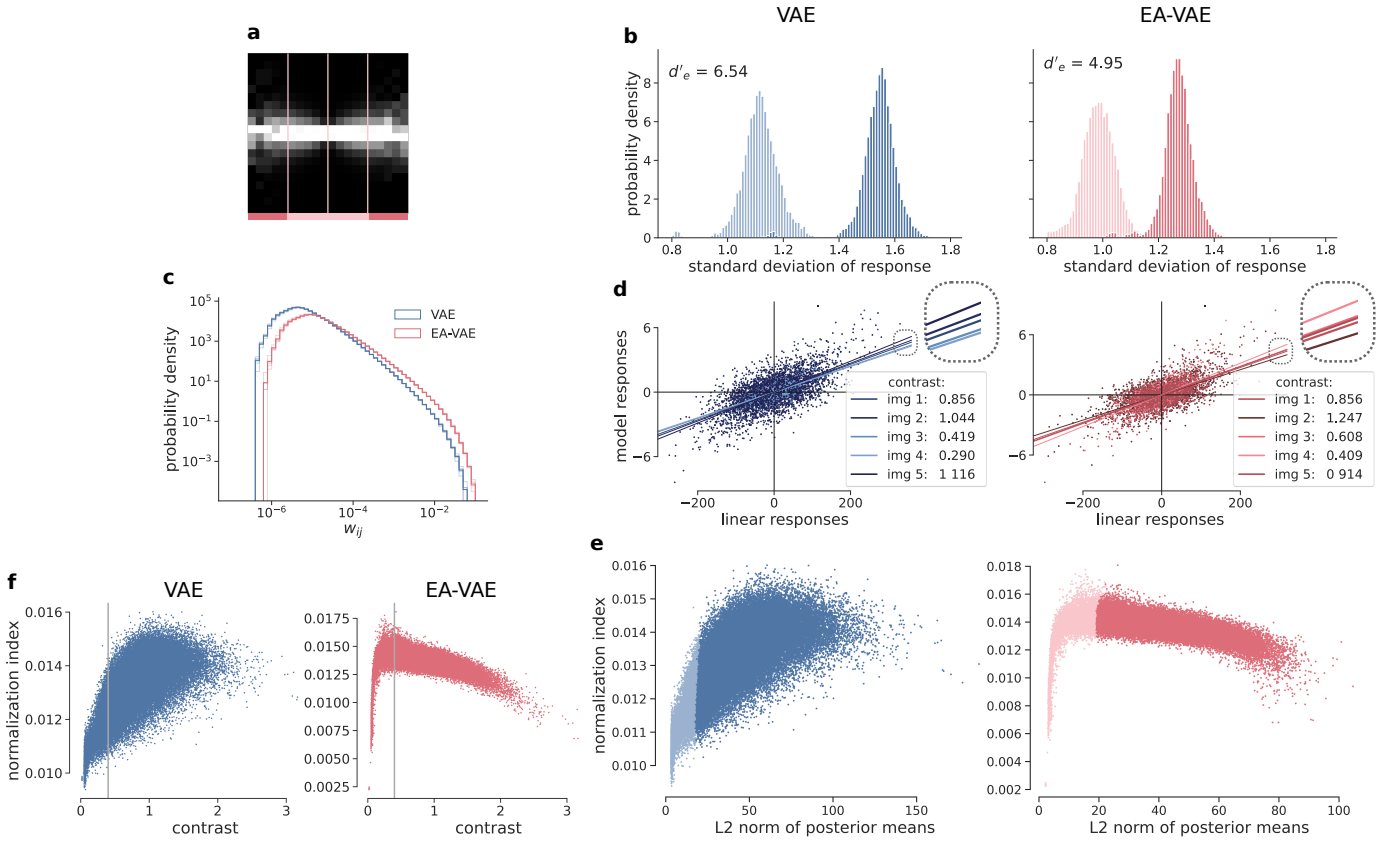


Fig. 7. Divisive normalization implemented by the recognition model. **a**, Conditional distribution ($p(L_1 | L_2, x)$) of linear filter responses of a pair of example latent variables over natural images. Linear filter responses are calculated as the dot product between an image and the receptive field of the neuron (Supplementary Section II-B3). L_1 and L_2 response intensities are shown on vertical and horizontal axes. Light and dark colored bars correspond to two central and two flanking quadrants of L_2 activation, respectively. **b**, Non-linear dependence of latent posteriors. Dependence is characterized by 100 randomly chosen pairs of latents using natural image posteriors. Standard deviation of the conditional distribution of posterior means are shown close to zero activation of the conditioned latent variable (two central quadrants) and at high-intensity activation (two flanking quadrants). **c**, Distribution of learned weights of the divisive normalization model for the standard VAE (blue) and EA-VAE (red) models. Dark line shows the average of five fits, individual fits are shown with light lines. The two distributions differ significantly (Kolmogorov-Smirnov test, $D = 0.598$, $p < .001$). **d**, Linear responses and posterior means in the standard VAE (blue) and EA-VAE (red) models of individual latents (dots) for a particular natural image. Colors distinguish responses to five different images, color lightness reflecting image contrast. Lines: linear fits, slope defines the normalization index (shown in the legend). **e**, Normalization index in the standard VAE (blue) and EA-VAE (red) models as a function of the normalization factor of the divisive normalization model, i.e. the Euclidean mean posterior means. Dots correspond to individual images, light colors correspond to images with contrast levels below observation noise (Supplementary Section II-B3). **f**, Same as **e** but normalization index is shown against image contrast. The vertical line: observation noise level.

These results confirm that differences in the properties of latent uncertainty representations between VAEs and EA-VAEs also translate to differences in output uncertainty, when those representations are used downstream for a supervised task.

G. Signatures of divisive normalization in the recognition model

So far, we have demonstrated that the inductive bias in the generative model contributes to more accurate inference. Next, we investigated how the recognition model is adapted to the inductive bias in the generative model. This ultimately promises to contribute to understanding the computational implementation that enables EA-VAEs to accomplish improved inference.

Contrast-related changes induce a dependency in the joint statistics of linear filters, referred to as bow-tie dependency: pairwise conditional distributions display a characteristic bow-tie shape (Fig. 7a). As VAEs assume independence of latent

variables, we draw inspiration from neuroscience, where divisive normalization was proposed to reclaim independence [25].

We trained the divisive normalization model (Eq. S39, Supplementary Section II-B3) on both the standard VAE and the EA-VAE. As the scaling variable explains contrast-related changes, we expected that the bow-tie dependency is reduced in EA-VAE. To analyze this, we computed the conditional distributions of the responses of latent pairs over natural images (Fig. 7a) and calculated the width of the conditional distribution both at low-, and high-level activations of the conditioned latent (Fig. 7b). High-level activations were characterized by higher variance in both the standard VAE and EA-VAE, but the difference was reduced in the EA-VAE (Fig. 7b). We found that the reduced bow-tie dependency could be explained by stronger divisive normalization in the EA-VAE, as reflected by stronger weights learned by the divisive normalization model (Fig. 7c).

We also assessed divisive normalization by measuring the

degree of suppression of individual latents depending on the activity of the remaining latent variables. Suppression was measured through the normalization index, which is based on the relationship between the linear responses and posterior means: a small normalization index indicated stronger suppression relative to the response expected from a linear model (Fig. 7d). We compared the response intensity of the latents with the normalization index for both the EA-VAE and standard VAE models (Fig. 7e). While the EA-VAE displayed consistently higher normalization with more intense latent activations, the standard VAE showed an opposite tendency, indicating that divisive normalization was characteristic of EA-VAE but not the standard VAE. Note that observation noise limits the quality of inference, as images with contrast lower than the level of observation noise are considered merely as noise by the generative model. Increased contrast results in higher signal mean (Fig. 1c-d) and therefore contrast might have a direct effect on the strength of divisive normalization. Indeed, contrast-dependence of the normalization index identified a strong effect of divisive normalization in the EA-VAE model, but not in the standard VAE (Fig. 7f).

In summary, we identified signatures of divisive normalization in EA-VAEs, supporting the conclusion that the inductive bias shapes the learning of the recognition model of the EA-VAE.

V. DISCUSSION

Deep generative models are praised for establishing a coherent framework for probabilistic machine learning [10], [45], and VAEs are particularly appealing as they provide a normative tool for learning of (and inference in) generative models of arbitrary complexity [33]. At the same time, the capacity of popular generative methods, and VAEs in particular, to faithfully represent the uncertainty of inferences has been put to question [22]. This makes the systematic study of uncertainty representations in VAEs key to know when not to trust them, and find a potential way forward to improve them. Indeed, our investigations consistently revealed pervasive problems in uncertainty representation of VAEs, which we argued stems from ignoring non-linear statistical dependencies characteristic of natural and artificial images. Motivated by earlier computer vision insights that addressed these non-linear dependencies through multiplicative latent variables as in the Gaussian Scale Mixture (GSM) model [24], we propose a novel framework that combines the simplicity and elegance of the VAE formulation with the flexibility of scale mixture models. In this model family, which we call EA-VAEs, the joint activation of a large number of local variables in the VAE can be *explained away* by an increase in the global multiplicative variable. This allows EA-VAEs to effectively capture non-linear dependencies in the data. Critically, we show that while the inclusion of this explaining away variable was initially motivated by the need to adequately model contrast-related changes in the uncertainty of inferences over natural images, the benefits of the framework extend to a variety of issues in a wide array of computer vision domains. This includes the highly relevant issues of uncertainty quantification when

faced with out-of-distribution, or corrupted data. Interestingly, our mathematical analyses of the EA-VAE model provided valuable insights into the geometrical properties of its learned representations, how these are fostered during training, and the corresponding mechanistic underpinnings of the solution. Indeed, our analyses revealed that along with the inductive bias introduced in the generative model, the recognition model of the VAE also underwent adaptation, developing signatures of divisive normalization, a widely reported motif in biological neural networks [31]. These results highlight the benefit of neuroscience-inspired ideas in artificial intelligence [46].

The proposed extension of the standard VAE model requires a small modification to the architecture, at no significant additional computational cost. We show however, that this simple intervention has a strong impact in the models' inductive biases, starkly improving on standard VAEs' capabilities. We argue that it is a principled extension because it directly addresses a common assumption in VAEs: the independence of latent variables. As argued, this assumption is easily violated. While this could in principle be addressed by learning a different representation to accommodate this dependency through strong nonlinearities in the generative model of VAEs, this does not occur in practice, where directly eliminating this source of nonlinearity appears to provide a more viable path. We have indeed investigated alternative strategies to remedy uncertainty representations that do not involve changes to the generative model. Specifically, we have explored MC dropout and IWAEs, which tackle two potential sources of uncertainty underestimation stemming from epistemic uncertainty and from the use of a single sample from the variational posterior to estimate the loss in VAEs. We found no evidence that either of these techniques could restore effective inference.

Moreover, we argue that the connection between the explaining away phenomenon in probabilistic inference and divisive normalization as a computational motif should be further explored and exploited in machine learning. As shown in Ref. [28], artificial neural systems optimized for full sampling-based Bayesian inference for models with explaining away variables result in solutions with mean responses which precisely recover divisive normalization. Our work provides evidence for this link in a variational setting, hinting at a wider pattern. It is hence possible that, as with divisive normalization in the brain, explaining away variables in artificial neural architectures may play an important role in multiple domains.

We have taken a thorough and layered approach to our study of EA-VAEs: building from first intuitions into the role of explaining away variables, which were then supported by analytical derivations and toy examples, escalating until reaching complex higher dimensional datasets. Throughout these cases we have systematically shown the beneficial properties of EA-VAEs. Inference models which can readily provide a meaningful representation of uncertainty are crucial in a plethora of fields, such as in the healthcare domain, where optimal information fusion relies on weighting multiple information sources by their relative uncertainties [47]. While further work is required to assess these benefits in other non-visual modalities, we believe EA-VAEs constitute a promising model type for a wide range of scenarios. Finally, the simplicity

and robustness of the proposed updated architecture promises effortless adoption of the EA-VAE architecture in practical applications.

VI. DATA AND CODE AVAILABILITY

All datasets used in this paper are publicly available. Van Hateren dataset of natural image patches is provided in Ref. [38]. MNIST and FashionMNIST datasets can be downloaded via Pytorch. ChestMNIST and NIH-ChestXray14 datasets can be downloaded and imported using the MEDM-NIST+ official repository. Inception pre-trained network weights can be imported with Pytorch, and ImageNet database can be downloaded from the official webpage.

Implementation code and instructions for reproducing results are freely available on GitHub, <https://github.com/josefina-catoni/EA-VAE-uncertainty-in-latent-representations>, <https://github.com/martosdomo/EA-VAE-natural-domain>.

REFERENCES

- [1] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Copyright Cambridge University Press, 2003.
- [2] E. T. Jaynes, *Probability Theory. The Logic of Science*, G. L. Bretthorst, Ed. Cambridge University Press: Cambridge, 2002.
- [3] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, and X. X. Zhu, "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, vol. 56, no. 1, pp. 1513–1589, Oct 2023. [Online]. Available: <https://doi.org/10.1007/s10462-023-10562-9>
- [4] J. O. Berger, *Statistical decision theory foundations, concepts, and methods* James O. Berger. Springer, 1980.
- [5] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.
- [6] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped dqn," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/8d8818c8e140c64c743113f563cf750f-Paper.pdf
- [7] N. Houlsby, F. Huszar, Z. Ghahramani, and M. Lengyel, "Bayesian active learning for classification and preference learning," 2011. [Online]. Available: <https://arxiv.org/abs/1112.5745>
- [8] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T.-Y. Liu, A. Manrai, D. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Veličković, M. Welling, L. Zhang, C. W. Coley, Y. Bengio, and M. Zitnik, "Scientific discovery in the age of artificial intelligence," *Nature*, vol. 620, no. 7972, pp. 47–60, Aug 2023. [Online]. Available: <https://doi.org/10.1038/s41586-023-06221-2>
- [9] J. Jiménez-Luna, F. Grisoni, and G. Schneider, "Drug discovery with explainable artificial intelligence," *Nature Machine Intelligence*, vol. 2, no. 10, pp. 573–584, Oct 2020. [Online]. Available: <https://doi.org/10.1038/s42256-020-00236-4>
- [10] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, pp. 7553, pp. 452–459, May 2015. [Online]. Available: <https://doi.org/10.1038/nature14541>
- [11] C. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics. Springer, 2006. [Online]. Available: <https://books.google.co.in/books?id=kTNoQgAACAAJ>
- [12] J. M. Tomczak, *Deep generative modeling*. Springer, 2022.
- [13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: <https://arxiv.org/abs/1406.2661>
- [14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022.
- [15] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," 2016. [Online]. Available: <https://arxiv.org/abs/1505.05770>
- [16] Z. Kadkhodaie and E. Simoncelli, "Stochastic solutions for linear inverse problems using the prior implicit in a denoiser," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 13 242–13 254. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/6e28943943dbed3c7f82fc05f269947a-Paper.pdf
- [17] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," 2021. [Online]. Available: <https://arxiv.org/abs/2011.13456>
- [18] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," 2015. [Online]. Available: <https://arxiv.org/abs/1503.03585>
- [19] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2017. [Online]. Available: <https://openreview.net/forum?id=Sy2fzU9gl>
- [20] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature Methods*, vol. 15, no. 12, p. 1053–1058, Nov 2018.
- [21] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf
- [22] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Do deep generative models know what they don't know?" *arXiv preprint arXiv:1810.09136*, 2018.
- [23] P. Berkes, G. Orbán, M. Lengyel, and J. Fiser, "Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment," *Science*, vol. 331, no. 6013, pp. 83–87, 2011.
- [24] M. J. Wainwright and E. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," in *Advances in Neural Information Processing Systems*, S.olla, T. Leen, and K. Müller, Eds., vol. 12. MIT Press, 1999. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/1999/file/6a5dfac4be1502501489f0f5a24b667-Paper.pdf
- [25] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control," *Nature neuroscience*, vol. 4, no. 8, pp. 819–825, 2001.
- [26] G. Orbán, P. Berkes, J. Fiser, and M. Lengyel, "Neural variability and sampling-based probabilistic representations in the visual cortex," *Neuron*, vol. 92, no. 2, pp. 530–543, 2016.
- [27] D. Festa, A. Aschner, A. Davila, A. Kohn, and R. Coen-Cagli, "Neuronal variability reflects probabilistic inference tuned to natural image statistics," *Nature Communications*, vol. 12, no. 1, p. 3635, Jun 2021. [Online]. Available: <https://doi.org/10.1038/s41467-021-23838-x>
- [28] R. Echeveste, L. Aitchison, G. Hennequin, and M. Lengyel, "Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference," *Nature neuroscience*, vol. 23, no. 9, pp. 1138–1149, 2020.
- [29] Y. LECUN, "The mnist database of handwritten digits," <http://yann.lecun.com/exdb/mnist/>, 1998. [Online]. Available: <https://cir.nii.ac.jp/crid/1571417126193283840>
- [30] X. Wang, Y. Peng *et al.*, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *CVPR*, 2017, pp. 3462–3471.
- [31] M. Carandini and D. Heeger, "Normalization as a canonical neural computation," *Nature Reviews Neuroscience*, vol. 13, no. 1, p. 51, 2012.
- [32] R. L. Goris, R. Coen-Cagli, K. D. Miller, N. J. Priebe, and M. Lengyel, "Response sub-additivity and variability quenching in visual cortex," *Nature Reviews Neuroscience*, vol. 25, no. 4, p. 237–252, Feb 2024.
- [33] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends in Machine Learning*: Vol. 12 (2019): No. 4, pp 307–392, 2019.
- [34] O. Rybkin, K. Daniilidis, and S. Levine, "Simple and effective vae training with calibrated decoders," 2021. [Online]. Available: <https://arxiv.org/abs/2006.13202>
- [35] V. Geadah, G. Barello, D. Greenidge, A. S. Charles, and J. W. Pillow, "Sparse-coding variational auto-encoders," *BioRxiv*, p. 399246, 2018.
- [36] F. Csikor, B. Meszéna, K. Ócsai, and G. Orbán, "Top-down perceptual inference shaping the activity of early visual cortex," *Nature Communications*, vol. 16, no. 1, p. 9998, 2025.
- [37] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*,

vol. 381, no. 6583, pp. 607–609, Jun 1996. [Online]. Available: <https://doi.org/10.1038/381607a0>

- [38] F. Csikor, B. Meszéna, B. Szabó, and G. Orbán, “Top-down inference in an early visual cortex inspired hierarchical variational autoencoder,” 2022.
- [39] J. H. van Hateren and A. van der Schaaf, “Independent component filters of natural images compared with simple cells in primary visual cortex,” *Proc Biol Sci*, vol. 265, no. 1394, pp. 359–366, Mar. 1998.
- [40] H. Xiao, K. Rasul, and R. Vollgraf, “Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms,” 2017. [Online]. Available: <https://arxiv.org/abs/1708.07747>
- [41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge,” *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [43] Y. Burda, R. Grosse, and R. Salakhutdinov, “Importance weighted autoencoders,” 2015. [Online]. Available: <https://arxiv.org/pdf/1509.00519>
- [44] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, “beta-vae: Learning basic visual concepts with a constrained variational framework,” *ICLR (Poster)*, vol. 3, 2017.
- [45] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [46] F. H. Sinz, X. Pitkow, J. Reimer, M. Bethge, and A. S. Tolias, “Engineering a less artificial intelligence,” *Neuron*, vol. 103, no. 6, pp. 967–979, Sep 2019. [Online]. Available: <https://doi.org/10.1016/j.neuron.2019.08.034>
- [47] M. A. Ricci Lara, C. Mosquera, E. Ferrante, and R. Echeveste, “Towards unraveling calibration biases in medical image analysis,” in *Workshop on Clinical Image-Based Procedures*. Springer, 2023, pp. 132–141.



Ferenc Csikor is a researcher at the Computational Systems Neuroscience Lab at HUN-REN Wigner Research Centre for Physics at Budapest, 1121, Hungary. His research interests include deep generative models and the normative modeling of mammalian vision. Csikor received his PhD in Physics from Eötvös University, which was followed by postdoctoral fellowships at The University of Edinburgh and at the Karlsruhe Institute of Technology.



Balázs Meszéna is a researcher at the Computational Systems Neuroscience Lab at HUN-REN Wigner Research Centre for Physics at Budapest, 1121, Hungary. His research interests include deep generative models, modeling biological vision and computational neuroscience in general. Meszéna received his PhD in Physics from Leiden University.



Enzo Ferrante is an Adjunct Researcher from Argentina’s National Research Council (CONICET) at the Institute of Computer Science (ICC) affiliated to Universidad de Buenos Aires and CONICET, at Ciudad Autónoma de Buenos Aires, 1428, Argentina. His research interests include machine learning for healthcare, robustness and fairness of AI. Ferrante received his PhD in Computer science from Université Paris-Saclay.



Josefina Catoni is a PhD Student in the specialty of Signals, Systems and Computational Intelligence at the Research Institute for Signals, Systems and Computational Intelligence - sinc(i), CONICET/UNL, Santa Fe, Santa Fe (3000), Argentina. Her research interests include computational neuroscience and representation of uncertainty in machine learning. Catoni received her MSc in Physics from the University of Buenos Aires. She is also a Chief of Practical Works at the National University of Litoral (UNL).



Diego H. Milone received the Bioengineering degree (Hons.) from National University of Entre Rios, Argentina, in 1998, and the Ph.D. degree in Microelectronics and Computer Architectures from Granada University, Spain, in 2003. He is Full Professor in the Department of Informatics at National University of Litoral (UNL). From 2009 to 2011 was Director of the Department of Informatics and from 2010 to 2014 was Assistant Dean for Science and Technology. Since 2006 he is a Research Scientist at the National Scientific and Technical Research Council (CONICET). He was founder and first Director of the Research Institute for Signals, Systems and Computational Intelligence, sinc(i), CONICET-UNL. His research interests include neural computing and statistical learning, with applications to biomedical signals and bioinformatics.



Domonkos Martos is a Research Assistant at Computational Systems Neuroscience Lab at HUN-REN Wigner Research Centre for Physics at Budapest, 1121, Hungary. His research interests lie at the intersection of mathematics, machine learning and sensory processing. Domonkos is pursuing MSc in Applied Mathematics at the Technical University, Budapest.



Gergő Orbán is Head of the Computational Systems Neuroscience Lab at HUN-REN Wigner Research Centre for Physics at Budapest, 1121, Hungary. His research interests include understanding biological intelligence through machine learning and improving machine intelligence through neuroscience. He has extensively contributed to normative models of vision and higher level cognition. Orbán received his PhD in biophysics from Eötvös University, which was followed by postdoctoral fellowships at Brandeis University and University of Cambridge.



Rodrigo Echeveste is Head of the Research Institute for Signals, Systems and Computational Intelligence - sinc(i), dependent on Argentina's National Research Council (CONICET) and Argentina's National University of Litoral (UNL). He holds a permanent position as Adjunct Researcher from CONICET and is an Adjunct Professor at UNL. His research interests include the study of perception and neural dynamics employing machine learning tools, and algorithmic fairness in deep learning. Echeveste received his BSc and MSc in Physics from Balseiro

Institute in Argentina, and his Phd in Natural Sciences with orientation Physics from the Goethe University in Frankfurt, Germany, followed by a PostDoc at the University of Cambridge, UK.