Aleatoric and Epistemic Uncertainty in Medical Imaging

A deep dive into the decomposition and its failure modes

Christian F. Baumgartner

Summer School for Uncertainty in ML 2026

Introduction

What we will cover today

This lecture is about the aleatoric-epistemic decomposition, and how it often breaks in real applications.

1. Very brief **recap of mathematical tools** required for this lecture.
2. **The decomposition.** Splitting predictive uncertainty into an *aleatoric* and an *epistemic* part, and how to measure each.
3. **Brief overview** of common approaches for estimating aleatoric and epistemic uncertainty.
4. **Discussion of benchmark studies** (classification and segmentation) showing the split often does *not* do what the theory promises.
5. **Discussion of the limitations** of the commonly used AU/EU split, and practical recommendations.

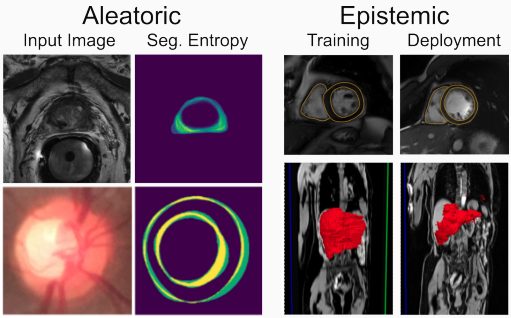
Lecture 2 will then address how we get clinical value out of uncertainty anyway.

Why do we care why a model is uncertain?

Knowing that a model is uncertain may not be enough. Sometimes we want to know *why*.

- Is the **data** inherently ambiguous? (e.g. a lesion boundary two radiologists would draw differently)
- Or has the **model** simply not seen enough similar cases? (e.g. a scanner or protocol outside the training distribution)

These call for *different actions*: defer to a human vs. acquire more data / flag out-of-distribution.



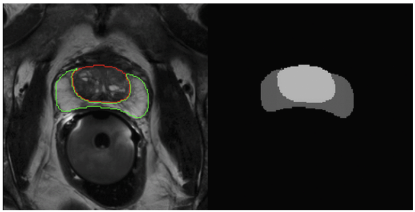
Why do we need the distinction between epistemic and aleatoric?

An example: Computing PSA density (PSAD) for prostate screening.

PSAD = PSA / Prostate Volume

Uncertainty in volume leads to uncertainty in PSAD.

- If segmentation uncertainty aleatoric
  → Report variance to user.
- If segmentation uncertainty epistemic
  → We cannot trust this prediction! Defer to human.



Mathematical Tools

Integrals, expectations, averages

An integral against a density is a weighted average

$$\int f(\theta)\, p(\theta)\, d\theta \;=\; \mathbb{E}_{p(\theta)}\big[\, f(\theta)\,\big]$$

The density p says *how much weight* each θ gets.

We can estimate expectations using Monte Carlo approximation

$$\mathbb{E}_{p(\theta)}\big[\, f(\theta)\,\big] \;\approx\; \frac{1}{M}\sum_{m=1}^{M} f(\theta_m), \qquad \theta_m \sim p(\theta)$$

Likely values of θ simply *show up more often* among the samples.

Variational inference

Given observations x and a latent variable z, the posterior $p(z \mid x) = \frac{p(x|z)\,p(z)}{p(x)}$ needs the normaliser $p(x) = \int p(x \mid z)\,p(z)\,dz$. This integral is typically *intractable*, so we cannot evaluate the posterior directly.

The idea: turn inference into optimisation

Give up on the exact posterior. Pick a family of *convenient* distributions $q_\phi(z)$ (e.g. diagonal Gaussians), then find its member closest to the truth:

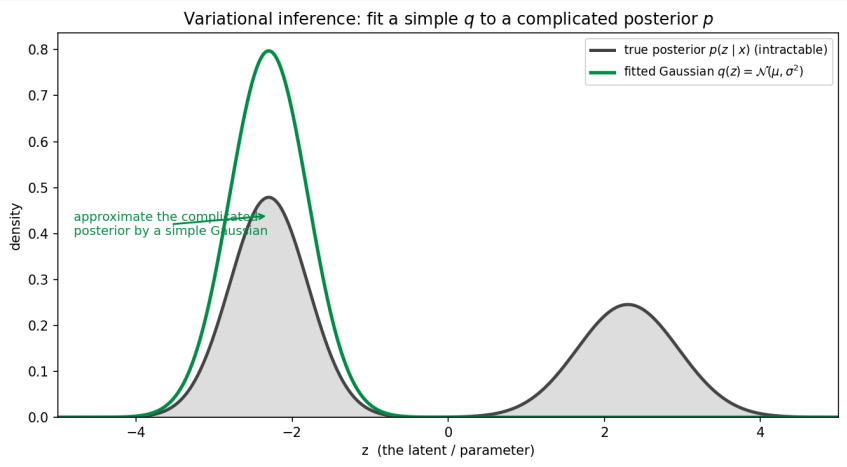
$$q_\phi = \arg\min_{\phi} \; \mathrm{KL}\big[\, q_\phi(z) \;\big\|\; p(z \mid x) \,\big]$$

KL is a mismatch score between two distributions: never negative, zero only if identical.

$p(x)$ does not depend on ϕ, so it *drops out*. What is left to maximise is the **ELBO**:

$$\underbrace{\mathbb{E}_{q_\phi}\big[\log p(x \mid z)\big]}_{\text{explain the data}} \;-\; \underbrace{\mathrm{KL}\big[q_\phi(z) \,\|\, p(z)\big]}_{\text{stay near the prior}}$$

Variational Inference



The Decomposition

Two kinds of uncertainty

Aleatoric (data) uncertainty

Irreducible noise / ambiguity in the data-generating process. Does *not* vanish with more data.

Epistemic (model) uncertainty

Uncertainty about the *model* (parameters, structure). *Reducible*: shrinks as we observe more data.

A cornerstone of the uncertainty quantification literature in recent years has been the claim that aleatoric and epistemic uncertainty can be separated.

Kendall & Gal, NeurIPS 2017 Hüllermeier & Waegeman, Mach. Learn. 2021

The Bayesian picture

For a prediction $y^\star$ at test input $x^\star$, with training data $\mathcal{D}$ and parameters $\boldsymbol{W}$, we can define:

- **Parameter uncertainty:** a posterior over weights $q(\boldsymbol{W}) \approx p(\boldsymbol{W} \mid \mathcal{D})$
  $\rightarrow$ **epistemic** uncertainty
- **Conditional noise:** the predictive $p(y^\star \mid \boldsymbol{W}, x^\star)$ at *fixed* weights
  $\rightarrow$ **aleatoric** uncertainty

The predictive distribution marginalises the weights out (the *Bayesian model average*):

$$p(y^\star \mid x^\star, \mathcal{D}) \;=\; \int p(y^\star \mid \boldsymbol{W}, x^\star)\, q(\boldsymbol{W})\, d\boldsymbol{W} \;=\; \mathbb{E}_{q(\boldsymbol{W})}[\, p(y^\star \mid \boldsymbol{W}, x^\star)\,] \,.$$

Strictly this is an *approximate* BMA: we integrate against q, not against the true posterior.

Depeweg et al., ICML 2018

How do we get $q(\boldsymbol{W})$? : epistemic methods

$p(\boldsymbol{W} \mid \mathcal{D})$ is intractable: the normaliser is an integral over millions of dimensions. All methods below are approximations that allow sampling.

Explicit approximations (a density we can write down)

- **Variational inference:** fit $q_\phi(\boldsymbol{W})$ (e.g. factorised Gaussian) by minimising $\mathrm{KL}\big[q_\phi \,\big\|\, p(\boldsymbol{W} \mid \mathcal{D})\big]$.
- **Laplace approximation:** Gaussian at a MAP mode, covariance from the local curvature (Hessian).
- **SWAG:** Gaussian fitted to the weights collected along the SGD trajectory.

Implicit approximations (defined by how we sample)

- **Deep ensembles:** M networks trained from different random initialisations; the members *are* the samples (different posterior modes).
- **MC dropout:** keep dropout active at test time; each random mask is a sample.

How do we get $p(y^\star \mid \boldsymbol{W}, x^\star)$? : aleatoric models

Closed form $\rightarrow$ no sampling needed

A standard classification / segmentation head: $p(y^\star \mid \boldsymbol{W}, x^\star) = \mathrm{Cat}\big(\mathrm{softmax}(f_{\boldsymbol{W}}(x^\star))\big)$.

At fixed $\boldsymbol{W}$ this is already a distribution. Uncertainty can (e.g.) be quantified with $H[\,p\,]$ (More on entropy will come on the next slide)

Generative $\rightarrow$ sample an inner latent (especially for segmentation)

$p(y^\star \mid \boldsymbol{W}, x^\star) = \int p(y^\star \mid \boldsymbol{W}, x^\star, z)\, p(z)\, dz$, estimated with K samples $z_k \sim p(z)$:

- **SSN:** z is Gaussian logit noise (low-rank spatial covariance).
- **Prob. U-Net / PHiSeg:** z is a cVAE latent (hierarchical in PHiSeg).
- **Diffusion:** z is the noise seed of the reverse process.
- **TTA:** z is the random augmentation applied at test time.

A quick primer: entropy and mutual information

Shannon entropy

$$H[p] \;:=\; -\sum_{a} p(a) \log p(a)$$

The entropy *of a distribution* p is a scalar measure of how "spread out" the distribution is.

Mutual information

$$I(a;b) \;=\; D_{\mathrm{KL}}\big[\, p(a,b) \,\big\|\, p(a)\,p(b) \,\big] \;=\; H[p(a)] \;-\; H[p(a \mid b)],$$

where $H[p(a \mid b)] = \mathbb{E}_{p(b)}\big[\, H[\, p(a \mid b)\,]\,\big]$ is the conditional entropy.

Intuition: the expected reduction in uncertainty about a from observing b (equivalently, about b from observing a).

Cover & Thomas 2006

The decomposition via entropy

The total uncertainty is the entropy of the predictive distribution:

$$\underbrace{H\big[\, p(y^\star \mid x^\star) \,\big]}_{\text{Total (TU)}} \;=\; H\big[\, \mathbb{E}_{q(\boldsymbol{W})}\, p(y^\star \mid \boldsymbol{W}, x^\star) \,\big].$$

By the definition of mutual information between $y^\star$ and $\boldsymbol{W}$ (everything conditional on $x^\star$),

$$\begin{aligned} I(y^\star; \boldsymbol{W} \mid x^\star) &= H\big[\, p(y^\star \mid x^\star) \,\big] - \mathbb{E}_{q(\boldsymbol{W})}\big[\, H[\, p(y^\star \mid \boldsymbol{W}, x^\star) \,]\,\big] \\ &= H\big[\, \mathbb{E}_{q(\boldsymbol{W})}\, p(y^\star \mid \boldsymbol{W}, x^\star) \,\big] - \mathbb{E}_{q(\boldsymbol{W})}\big[\, H[\, p(y^\star \mid \boldsymbol{W}, x^\star) \,]\,\big]. \end{aligned}$$

The decomposition via entropy

The total uncertainty is the entropy of the predictive distribution:

$$\underbrace{H\big[\, p(y^\star \mid x^\star) \,\big]}_{\text{Total (TU)}} \;=\; H\big[\, \mathbb{E}_{q(\boldsymbol{W})}\, p(y^\star \mid \boldsymbol{W}, x^\star) \,\big].$$

By the definition of mutual information between $y^\star$ and $\boldsymbol{W}$ (everything conditional on $x^\star$),

$$\begin{aligned} I(y^\star; \boldsymbol{W} \mid x^\star) &= H\big[\, p(y^\star \mid x^\star) \,\big] - \mathbb{E}_{q(\boldsymbol{W})}\big[\, H[\, p(y^\star \mid \boldsymbol{W}, x^\star) \,]\,\big] \\ &= H\big[\, \mathbb{E}_{q(\boldsymbol{W})}\, p(y^\star \mid \boldsymbol{W}, x^\star) \,\big] - \mathbb{E}_{q(\boldsymbol{W})}\big[\, H[\, p(y^\star \mid \boldsymbol{W}, x^\star) \,]\,\big]. \end{aligned}$$

Rearranging gives the additive decomposition:

$$\underbrace{H\big[\, \mathbb{E}_q\, p \,\big]}_{\text{TU}} \;=\; \underbrace{\mathbb{E}_q\, H[\, p \,]}_{\text{Aleatoric (AU)}} \;+\; \underbrace{I(y^\star; \boldsymbol{W} \mid x^\star)}_{\text{Epistemic (EU)}}$$

Depeweg et al., ICML 2018 Kendall & Gal, NeurIPS 2017

Why is $\mathbb{E}_q H[p]$ the aleatoric part?

Step 1: conditioning on W removes epistemic uncertainty

Fixing $W = W_m$ leaves nothing unknown about the model, so there is no epistemic uncertainty.

Any spread left in $p(y^\star \mid W_m, x^\star)$ therefore can only come from the data.

$$H\big[\, p(y^\star \mid W_m, x^\star) \,\big] \;=\; \text{aleatoric uncertainty } \textit{of that model}$$

Step 2: we do not know which W is correct

So we average each model's aleatoric estimate over the posterior:
$\mathrm{AU} = \mathbb{E}_{q(W)}\big[\, H[\, p(y^\star \mid W, x^\star)\,]\,\big]$.

Sanity check: with infinite data $q(W) \to \delta_{W^*}$, so EU $\to 0$ but AU remains.

Why is $I(y^\star; W)$ the epistemic part?

Mutual information is symmetric:

$$I(y^\star; \boldsymbol{W} \mid x^\star) = \underbrace{H[y^\star] - H[y^\star \mid \boldsymbol{W}]}_{\text{what we learn about } y^\star \text{ from } \boldsymbol{W}} = \underbrace{H[\boldsymbol{W}] - H[\boldsymbol{W} \mid y^\star]}_{\text{what we learn about } \boldsymbol{W} \text{ from } y^\star}$$

- $I(y^\star; \boldsymbol{W} \mid x^\star) = 0 \iff y^\star \perp \boldsymbol{W}$: all models under $q(\boldsymbol{W})$ predict the *same* thing $\Rightarrow$ no epistemic uncertainty.

Connection to Bayesian Active Learning by Disagreement (BALD)

In the active learning method BALD, $I(y^\star; \boldsymbol{W} \mid x^\star)$ is used as the *expected information gain about the parameters* from observing the label of $x^\star$. Label the point that teaches you most about $\boldsymbol{W}$.

Houlsby et al. 2011 (BALD) Gal, Islam, Ghahramani, ICML 2017

Entropy of the average vs. average of the entropies

Writing it down again:

$$\mathrm{TU} = H\big[\,\overline{p}\,\big], \qquad \overline{p} = \mathbb{E}_{q(\boldsymbol{W})}[\,p(y \mid \boldsymbol{W}, x)\,] \quad \textbf{entropy of the average}$$

$$\mathrm{AU} = \mathbb{E}_{q(\boldsymbol{W})}\big[\,H[\,p(y \mid \boldsymbol{W}, x)\,]\,\big] \quad \textbf{average of the entropies}$$

$$\mathrm{EU} = \mathrm{TU} - \mathrm{AU} \;=\; I(y; \boldsymbol{W} \mid x) \;\geq\; 0$$

Sampling estimator with $\boldsymbol{W}_m \sim q(\boldsymbol{W})$, $\pi_m = p(y \mid \boldsymbol{W}_m, x)$:

$$\mathrm{TU} = H\Big[\tfrac{1}{M}\textstyle\sum_m \pi_m\Big], \qquad \mathrm{AU} = \tfrac{1}{M}\textstyle\sum_m H[\pi_m], \qquad \mathrm{EU} = \mathrm{TU} - \mathrm{AU}.$$

Important Note

EU is defined as a *residual*, $\mathrm{EU} = \mathrm{TU} - \mathrm{AU}$. It is not measured independently.

The same story with variances

For real-valued outputs, the **law of total variance** gives the identical structure:

$$\underbrace{\mathrm{Var}(y^\star \mid x^\star)}_{\text{Total}} = \underbrace{\mathbb{E}_{q(\boldsymbol{W})}\big[\mathrm{Var}(y^\star \mid \boldsymbol{W}, x^\star)\big]}_{\text{Aleatoric: mean of the variances}} + \underbrace{\mathrm{Var}_{q(\boldsymbol{W})}\big(\mathbb{E}[y^\star \mid \boldsymbol{W}, x^\star]\big)}_{\text{Epistemic: variance of the means}} .$$

Same procedure, just with variance instead of entropy. Neither is the *more correct* one, but the entropy-based decomposition is more widespread.

Depeweg et al., ICML 2018

No separate aleatoric model is required

A cool property of the information-theoretic decomposition is that you need **only a distribution over weights**. No separate aleatoric model is required.

The recipe

1. Draw M weight samples $\boldsymbol{W}_m \sim q(\boldsymbol{W})$.
2. Predict: $\pi_m = p(y^\star \mid \boldsymbol{W}_m, x^\star)$ (for classification, just the softmax output).
3. Read off all three numbers from the *same* set $\{\pi_m\}$:

$$\mathrm{TU} = H\Big[\tfrac{1}{M}\textstyle\sum_m \pi_m\Big], \qquad \mathrm{AU} = \tfrac{1}{M}\textstyle\sum_m H[\pi_m], \qquad \mathrm{EU} = \mathrm{TU} - \mathrm{AU}.$$

Example. A **deep ensemble of 5 plain U-Nets / ResNets**, trained with nothing but cross-entropy, already yields AU, EU and TU.

Depeweg et al., ICML 2018 Lakshminarayanan et al., NeurIPS 2017 Gal & Ghahramani, ICML 2016

Quick overview of common EU and AU estimation approaches

Epistemic methods: deep ensembles

Idea. Train M (typically 5–10) networks independently, from different random initialisations (and data shuffling). Each settles in a different basin of the loss landscape.

- The trained weights $\{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_M\}$ are treated as samples from $q(\boldsymbol{W})$: an *implicit* posterior.
- Predict by averaging: $\overline{p} = \frac{1}{M} \sum_m p(y^\star \mid \boldsymbol{W}_m, x^\star)$.

Trade-offs

+ Simple; consistently the *best* performing and best calibrated in practice; genuinely diverse (different modes).
– $M\times$ **training** and inference cost.

Lakshminarayanan, Pritzel, Blundell, NeurIPS 2017

Epistemic methods: MC dropout

Idea. Dropout randomly zeroes units during training (regularisation). *Keep it on at test time.*

- Each stochastic forward pass uses a different random mask $\Rightarrow$ a different sub-network $\Rightarrow$ a sample $\boldsymbol{W}_t \sim q(\boldsymbol{W})$.
- T forward passes give T predictions; average them / take their entropy.
- Gal & Ghahramani: this approximates variational inference, where the dropout masks are Bernoulli and the induced $q(\boldsymbol{W})$ is a mixture of point masses.

Trade-offs

+ Straightforward training; cheap samples (just multiple forward passes); no architectural changes required.

– Limited diversity; heavily criticised theoretically (q is a mixture of deltas, which is a poor posterior model and makes the KL term ill-behaved).

Epistemic methods: stochastic weight averaging-Gaussian (SWAG)

Idea. Towards the end of training with SGD, the iterates bounce around a mode. We can use these weight samples to fit a distribution.

- Collect weight snapshots from the last epochs and fit a Gaussian:

$$q(\boldsymbol{W}) = \mathcal{N}\big(\boldsymbol{W}_{\text{SWA}},\ \tfrac{1}{2}(\boldsymbol{\Sigma}_{\text{diag}} + \boldsymbol{\Sigma}_{\text{low-rank}})\big),$$

- Sample weights from this Gaussian at test time. **SWAG-D** = diagonal covariance only (cheaper).

Trade-offs

+ Cheap add-on to ordinary SGD training; captures local posterior *geometry*.

– Single mode (local), so less diverse than ensembles.

+/– Works with other optimisers (e.g. Adam), though the SGD-as-approximate-sampler justification no longer applies.

Maddox, Garipov, Izmailov, Vetrov, Wilson, NeurIPS 2019

Epistemic methods: Laplace approximation

Idea. Approximate the posterior around a trained (MAP) solution $\boldsymbol{W}_{\text{MAP}}$ by a Gaussian, from a 2nd-order Taylor expansion of the log-posterior:

$$q(\boldsymbol{W}) = \mathcal{N}\big(\boldsymbol{W}_{\text{MAP}},\ \boldsymbol{H}^{-1}\big), \qquad \boldsymbol{H} = -\nabla^2 \log p(\boldsymbol{W} \mid \mathcal{D})\big|_{\boldsymbol{W}_{\text{MAP}}}.$$

- $\boldsymbol{H}$ (the Hessian) is approximated by the generalised Gauss–Newton / Fisher matrix, which is guaranteed positive semi-definite. Even that is intractable for whole networks, so further structural approximations (diagonal, KFAC) are used.
- **Last-layer Laplace** is a cheap and popular alternative.

Trade-offs

+ Principled; can be applied post-hoc; very cheap in the last-layer form.
– Local Gaussian at a single mode: assumes the loss is locally quadratic (deep nets are not); the structural approximations discard real parameter correlations.

MacKay 1992 Daxberger et al., NeurIPS 2021 (Laplace Redux)

Laplace approximation: Graphical intuition

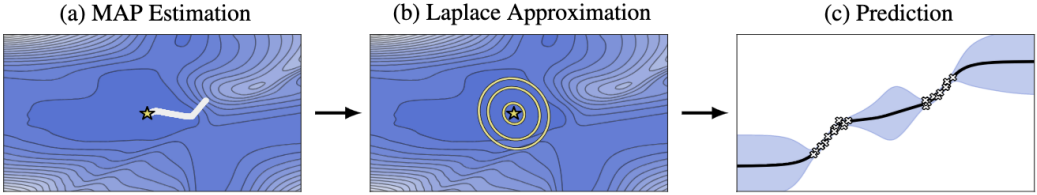
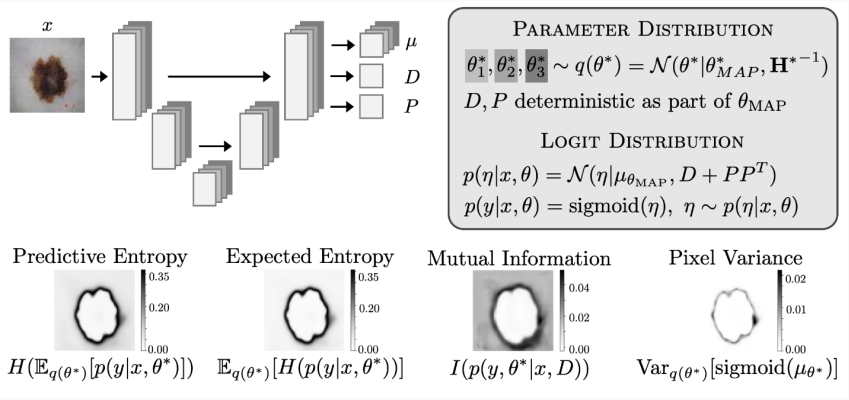


Figure 1: Probabilistic predictions with the Laplace approximation in three steps. (a) We find a MAP estimate (yellow star) via standard training (background contours = log-posterior landscape on the two-dimensional PCA subspace of the SGD trajectory [30]). **(b)** We locally approximate the posterior landscape by fitting a Gaussian centered at the MAP estimate (yellow contours), with covariance matrix equal to the negative inverse Hessian of the loss at the MAP—this is the Laplace approximation (LA). **(c)** We use the LA to make predictions with *predictive uncertainty estimates*—here, the black curve is the predictive mean, and the shading covers the 95% confidence interval.

(Screenshot from Paper).

Laplace approximation for U-Nets



Quick shoutout to former DTU PhD student Kilian Zepf who extended Laplace Approximations to segmentation UNets.

Epistemic methods: Bayesian neural networks

Idea. Place a prior $p(\boldsymbol{W})$; the posterior is $p(\boldsymbol{W} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{W})\, p(\boldsymbol{W})$. Approximate it with a variational distribution q_ϕ.

Variational inference (e.g. Bayes by Backprop)

Fit $q_\phi(\boldsymbol{W})$ (e.g. mean-field Gaussian) by maximising the ELBO:

$$\max_{\phi} \; \mathbb{E}_{q_\phi}\big[\log p(\mathcal{D} \mid \boldsymbol{W})\big] - \mathrm{KL}\big[q_\phi(\boldsymbol{W}) \,\big\|\, p(\boldsymbol{W})\big].$$

Trained end-to-end with the reparameterisation trick.

Trade-offs

+ Principled theoretical framework.

– Hard to scale; requires extra GPU memory for training (a mean and a variance per weight); in practice often worse than ensembles.

MacKay 1992 Neal 1996 Blundell et al., ICML 2015 (Bayes by Backprop)

Bayesian Neural Netowrk: Graphical intuition

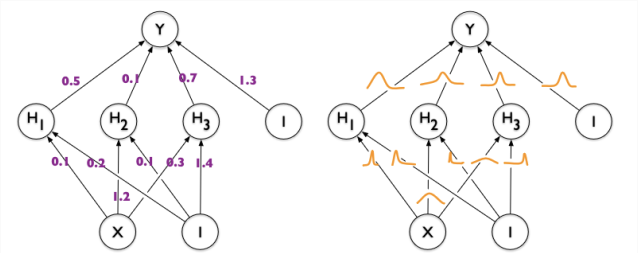


Figure 1. Left: each weight has a fixed value, as provided by classical backpropagation. Right: each weight is assigned a distribution, as provided by Bayes by Backprop.

(Screenshot from Paper).

Blundell et al., ICML 2015 (Bayes by Backprop)

Epistemic methods: deterministic (distance-aware)

Idea. Measure *distance from the training data* in feature space (an input unlike anything in the training set is epistemically uncertain). No sampling from $q(\boldsymbol{W})$ required.

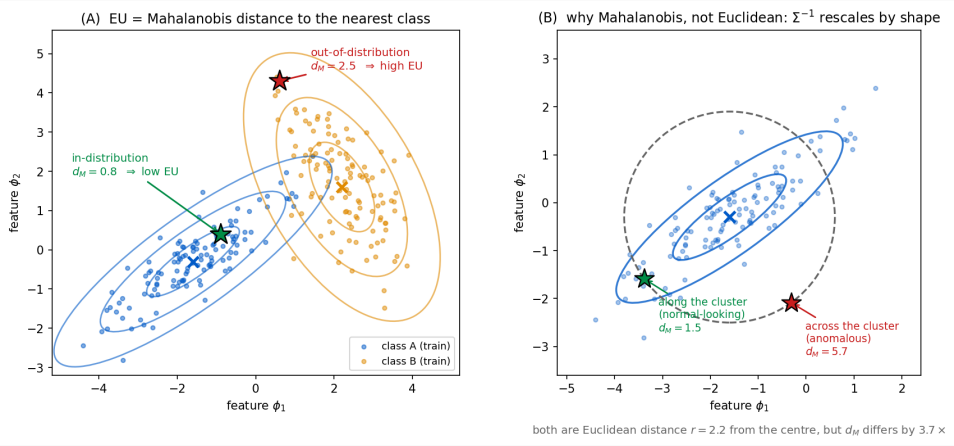
- **Feature-density (Mahalanobis / DDU):** fit class-conditional Gaussians / GMM to the penultimate features $\phi(x)$ with a *shared* (tied) covariance $\boldsymbol{\Sigma}$;
  $\text{EU} = \min_c (\phi(x) - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}^{-1} (\phi(x) - \boldsymbol{\mu}_c)$.
- **Learned distance (DUQ):** RBF distance to learned class centroids + gradient penalty.

Trade-offs

+ One pass; cheap.
– Needs feature regularisation (spectral norm) to avoid feature *collapse*; gives no principled AU/EU decomposition (the score is not on the same scale as an entropy).

Lee et al. 2018 Mukhoti et al. 2023 (DDU) Van Amersfoort et al. 2020 (DUQ) Liu et al. 2020 (SNGP)

Epistemic methods: deterministic (distance-aware)



The aleatoric part: softmax and cross-entropy

For the aleatoric component the **softmax output is very often used directly**: its entropy $H[p(y^\star \mid \boldsymbol{W}, x^\star)]$ *is* the per-model aleatoric estimate.

Cross-entropy is a likelihood

Training with cross-entropy = maximum likelihood under a categorical model (a Bernoulli in the binary case). For a one-hot label y:

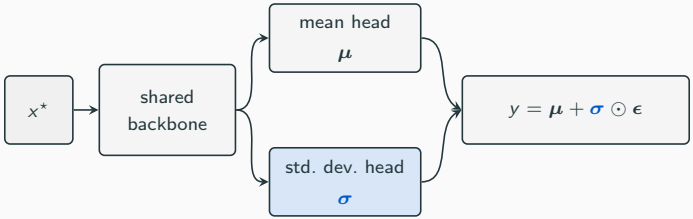
$$\mathcal{L}_{\mathrm{CE}} = -\sum_c y_c \log p_c = -\log \mathrm{Cat}(y \mid \boldsymbol{p})$$

Toy check

If one input x appears in training with label 0 half the time and 1 half the time, the expected loss $-\frac{1}{2}\log p - \frac{1}{2}\log(1-p)$ is minimised at $p = 0.5$: the network learns the true label noise.

Modelling aleatoric noise explicitly (regression)

Another strategy is to use two heads on a common backbone: one for the mean $\boldsymbol{\mu}$ and one for the standard deviation $\boldsymbol{\sigma}$ (the aleatoric uncertainty) of a Gaussian.



with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. **Heteroscedastic (Gaussian NLL) loss:**

$$\mathcal{L} = \frac{1}{N} \sum_i \left[\frac{\left(y_i - \boldsymbol{\mu}(x_i)\right)^2}{2\,\boldsymbol{\sigma}(x_i)^2} + \tfrac{1}{2} \log \boldsymbol{\sigma}(x_i)^2 \right]$$

The net can *down-weight* noisy points via a large $\boldsymbol{\sigma}$, but pays a $\frac{1}{2} \log \boldsymbol{\sigma}^2$ penalty for it.

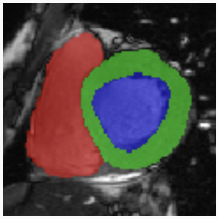
Kendall & Gal, NeurIPS 2017 Nix & Weigend 1994

Aleatoric models for pixel-wise prediction (e.g. segmentation)

In segmentation the output is an entire label *map* $y \in \{1, \dots, C\}^{H \times W}$. The aleatoric distribution is a distribution over *whole maps*.

The problem with per-pixel softmax

A softmax gives *independent* per-pixel marginals $p(y_i \mid x)$. Sampling each pixel independently yields *spatially incoherent* noise (salt-and-pepper), not plausible segmentation maps.



- Pixel values of a segmentation map are **spatially correlated**: a boundary shifts coherently, or raters disagree about *whole structures*.
- So we need to model the **joint** $p(y \mid \boldsymbol{W}, x)$ with its pixel *covariance*.

Kohl et al. 2018; Monteiro et al. 2020; Baumgartner et al. 2019

Aleatoric methods: stochastic segmentation networks (SSN)

Problem. We cannot model the full covariance of a segmentation map (too big).
Idea. Model the *logits* as a low-rank-plus-diagonal multivariate Gaussian over all pixels:

$$\boldsymbol{\eta}(x) \sim \mathcal{N}\big(\boldsymbol{\mu}(x),\ \boldsymbol{\Sigma}(x)\big), \qquad \boldsymbol{\Sigma}(x) = \boldsymbol{D}(x) + \boldsymbol{P}(x)\boldsymbol{P}(x)^{\top}, \qquad y \sim \mathrm{Cat}\big(\mathrm{softmax}(\boldsymbol{\eta})\big).$$

- $\boldsymbol{D}$ diagonal (per-pixel noise) + $\boldsymbol{P}$ a **rank-r** factor ($r \approx 10$) capturing the dominant *spatial* covariance.
- Sample logits $\rightarrow$ softmax $\rightarrow$ a *coherent* label map. Trained by MC log-likelihood of the (multiple) annotations (logistic-normal).

Trade-offs

+ Explicit, interpretable covariance; cheap; coherent samples.
– Gaussian-in-logit family is restrictive; low rank caps complexity.

Monteiro et al., NeurIPS 2020

Aleatoric methods: probabilistic U-Net (cVAE)

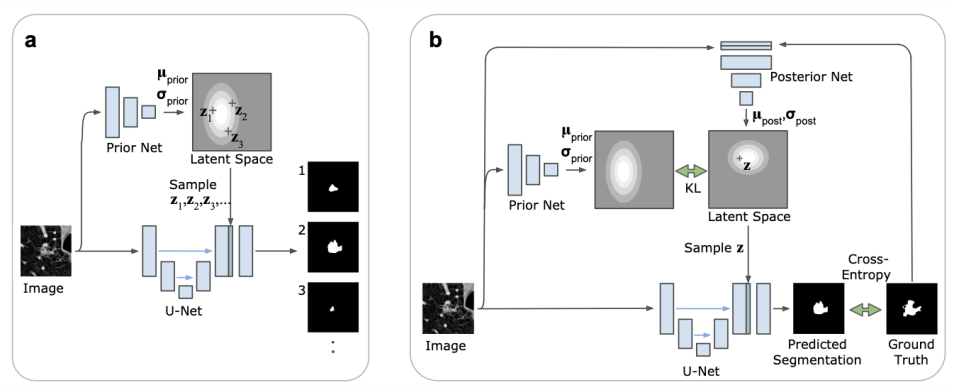
Idea. Model the variation using a low-dimensional latent random variable z that captures the structure of plausible segmentations (just like VAEs for generating images).
Note: the idea dates back to the cVAE (Sohn et al., NeurIPS 2015).

- A **prior net** gives $p(z \mid x) = \mathcal{N}\big(\boldsymbol{\mu}_{\mathrm{p}}(x),\, \mathrm{diag}(\boldsymbol{\sigma}^2_{\mathrm{p}}(x))\big)$ (low-dim, e.g. 6-D). A **posterior net** $q(z \mid x, y)$ is used only during training.
- A sample z is broadcast and concatenated with the U-Net feature map; a few 1×1 convs produce the mask. Different $z \Rightarrow$ different *coherent* segmentation.
- Trained as a cVAE: $\mathcal{L} = \mathbb{E}_{q(z|x,y)}[\mathrm{CE}] + \beta\, \mathrm{KL}\big[q(z \mid x, y) \,\|\, p(z \mid x)\big]$.

Trade-offs

+ Globally coherent, diverse samples; conceptually clean.
– One global low-dim latent, injected only at the end $\Rightarrow$ limited *multi-scale* / local variation; sample diversity can be modest.

Aleatoric methods: probabilistic U-Net (cVAE)



(screenshot from paper)

Kohl et al., NeurIPS 2018

Aleatoric methods: PHiSeg – a hierarchical cVAE

Idea. One global latent is too coarse. Use a *hierarchy* of latents $z_1, \ldots, z_L$, one per resolution scale.

- Coarse latents govern **global** structure (is the lesion there? where roughly?); fine latents govern **local** detail (exact boundary).
- Separate prior and posterior at *each* level. The mask is assembled across scales.

Trade-offs

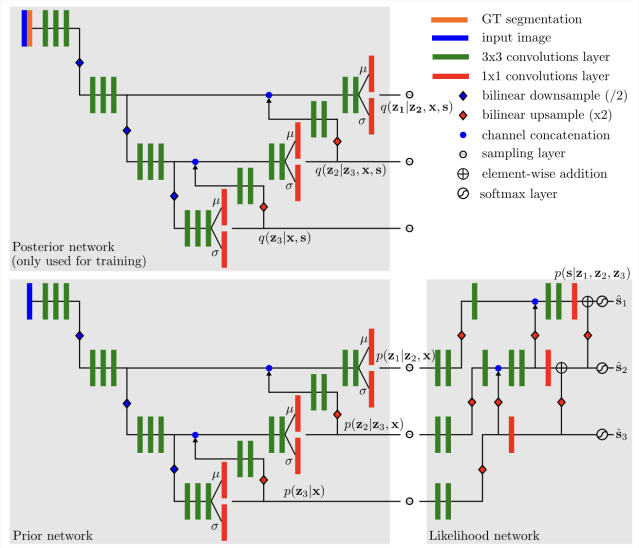
+ Captures *multi-scale* ambiguity; more diverse, better-calibrated samples than the Prob. U-Net; separates global from local variation.

– More complex; still a Gaussian-latent hierarchy; requires dedicated architecture.

Kohl et al.'s hierarchical Prob. U-Net (Kohl et al. 2019) is a parallel development with the same multi-scale-latent idea.

Baumgartner et al., MICCAI 2019 Kohl et al. 2019

Aleatoric methods: PHiSeg – a hierarchical cVAE



Aleatoric methods: diffusion models

Idea. Learn to *reverse a noising process* on the segmentation map, conditioned on the image x.

- Start from Gaussian noise; iteratively denoise to a mask, conditioned on x at every step.
- Different **noise seeds** $\Rightarrow$ different samples of $p(y \mid \boldsymbol{W}, x)$.
- Can represent *complex, multimodal* distributions over masks.

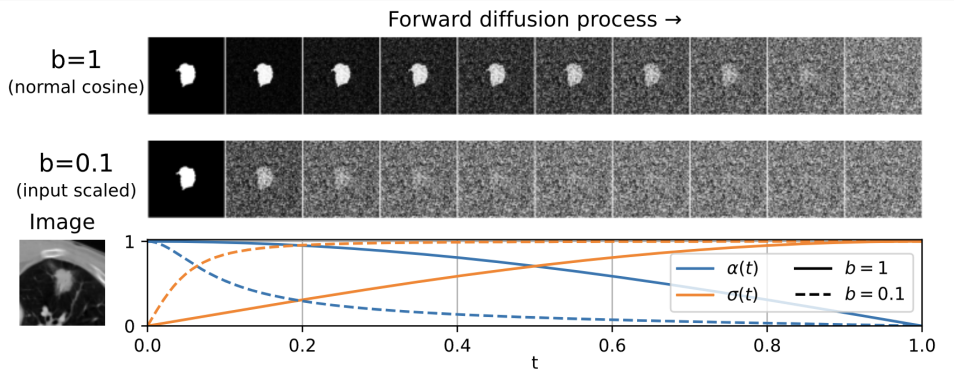
Trade-offs

+ Most *expressive* aleatoric method in this lecture.

– **Slow**: many denoising steps *per sample*, and UQ needs many samples $\Rightarrow$ expensive.

Amit et al. 2021 (SegDiff) Wolleb et al. 2022 Rahman et al., CVPR 2023

Aleatoric methods: diffusion models – A local example

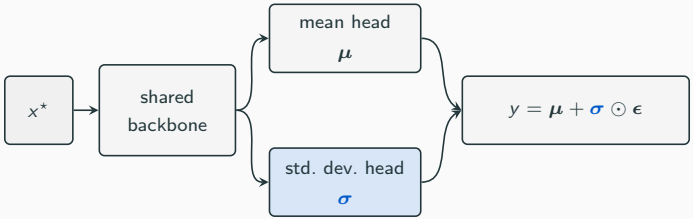


(screenshot from paper)

Christensen et al. Scandinavian Conference on Image Analysis 2025

Modelling EU and AU independently

Reminder: variance prediction for aleatoric uncertainty estimation



with $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. **Heteroscedastic (Gaussian NLL) loss:**

$$\mathcal{L} = \frac{1}{N} \sum_i \left[\frac{\big(y_i - \boldsymbol{\mu}(x_i)\big)^2}{2\,\boldsymbol{\sigma}(x_i)^2} + \tfrac{1}{2} \log \boldsymbol{\sigma}(x_i)^2 \right]$$

The net can *down-weight* noisy points via a large $\boldsymbol{\sigma}$, but pays a $\frac{1}{2} \log \boldsymbol{\sigma}^2$ penalty for it.

Kendall & Gal, NeurIPS 2017 Nix & Weigend 1994

Combining variance prediction with epistemic sampling (regression)

We can add a weight distribution $q(\boldsymbol{W})$ on top of the noise head and arrive at a split equivalent to the **law of total variance**:

$$\underbrace{\mathrm{Var}(y^\star \mid x^\star)}_{\text{Total}} = \underbrace{\mathbb{E}_{q(\boldsymbol{W})}\big[\boldsymbol{\sigma}^2(x^\star, \boldsymbol{W})\big]}_{\text{AU: mean of the learned noise}} + \underbrace{\mathrm{Var}_{q(\boldsymbol{W})}\big(\boldsymbol{\mu}(x^\star, \boldsymbol{W})\big)}_{\text{EU: spread of the means}}$$

Estimated with M weight samples $\boldsymbol{W}_m \sim q(\boldsymbol{W})$ (MC dropout, as in Kendall & Gal; or an ensemble):

$$\text{AU} \approx \tfrac{1}{M}\textstyle\sum_m \boldsymbol{\sigma}_m^2, \qquad \text{EU} \approx \tfrac{1}{M}\textstyle\sum_m \boldsymbol{\mu}_m^2 - \big(\tfrac{1}{M}\textstyle\sum_m \boldsymbol{\mu}_m\big)^2.$$

- This is exactly the variance decomposition from before, but AU is now a *learned* head.
- AU and EU are estimated **separately and simultaneously** : EU is no longer a residual.

Kendall & Gal, NeurIPS 2017

The classification version (Gaussian logits)

For classification a Gaussian is fit to the logits, which are then squashed through a softmax:

$$\boldsymbol{z}_t(x^\star) = \boldsymbol{\mu}(x^\star, \boldsymbol{W}) + \boldsymbol{\sigma}(x^\star, \boldsymbol{W}) \odot \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \quad \boldsymbol{\pi} = \tfrac{1}{T} \textstyle\sum_t \mathrm{softmax}(\boldsymbol{z}_t).$$

- Train by sampling T logit draws, averaging the softmax, and applying cross-entropy ("learned loss attenuation" for classification).
- The same $\boldsymbol{\sigma}$-head / $q(\boldsymbol{W})$ split applies.

The catch

The clean variance identity holds *in logit space*. After the **softmax nonlinearity** the additive AU + EU split is only *approximate*.

Kendall & Gal, NeurIPS 2017 Monteiro et al., NeurIPS 2020 (SSN)

Combining separate AU and EU mechanisms in the IT decomposition

We can also use both aleatoric and epistemic samples to feed the IT decomposition.

Nested sampling

Draw M weight samples $\boldsymbol{W}_m \sim q(\boldsymbol{W})$ (the EU mechanism, e.g. dropout / ensemble), and for each, K aleatoric samples z_k (e.g. PHiSeg latents / diffusion seeds):

$$\pi_{m,k} = p(y^\star \mid \boldsymbol{W}_m, x^\star, z_k), \qquad \bar{\pi}_m = \tfrac{1}{K} \textstyle\sum_k \pi_{m,k}.$$

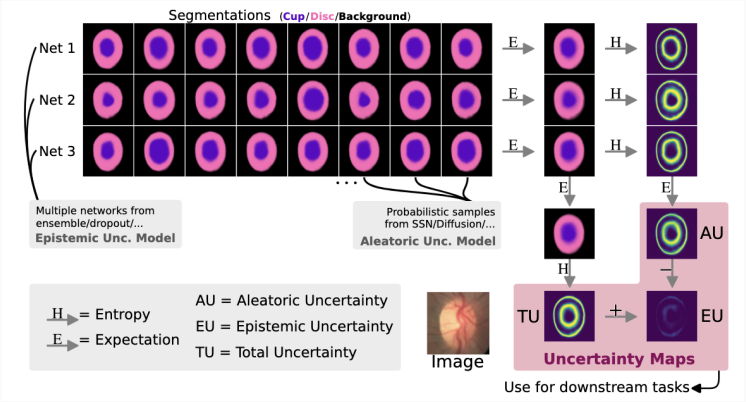
Then just apply the decomposition with the aleatoric average prediction $\bar{\pi}_m$:

$$\mathrm{AU} = \tfrac{1}{M} \textstyle\sum_m H[\bar{\pi}_m], \qquad \mathrm{TU} = H\big[\tfrac{1}{M} \textstyle\sum_m \bar{\pi}_m\big], \qquad \mathrm{EU} = \mathrm{TU} - \mathrm{AU}.$$

The AU mechanism supplies the *inner* distribution $p(y^\star \mid \boldsymbol{W}, x^\star)$; the EU mechanism supplies the *outer* average over $q(\boldsymbol{W})$.

Christensen et al. arxiv (2026)

Combining separate AU and EU mechanisms – Visual intuition

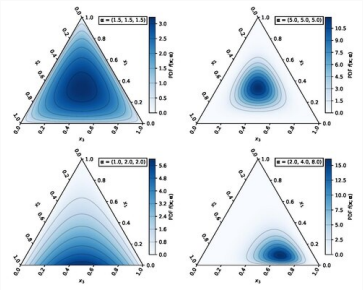


$$\mathrm{AU} = \frac{1}{M}\sum_m H[\bar{\pi}_m], \qquad \mathrm{TU} = H\Big[\frac{1}{M}\sum_m \bar{\pi}_m\Big], \qquad \mathrm{EU} = \mathrm{TU} - \mathrm{AU}.$$

Christensen et al. arxiv (2026)

Yet another idea: evidential deep learning (EDL)

Ensembles and MC dropout give a *cloud of points* on the probability simplex: one predicted $\boldsymbol{p}$ per sampled $\boldsymbol{W}_m$. EDL instead estimates a **density over that simplex** directly by fitting a **distribution over distributions**. The network outputs non-negative **evidence** per class, parameterising a Dirichlet.



The 2-simplex for $K = 3$

A Dirichlet over the simplex

- **Aleatoric**: *where the blob sits*. Near the centre ⇒ the classes overlap.
- **Epistemic**: *how spread out* it is. Little evidence ⇒ diffuse blob ⇒ uncertain $\boldsymbol{p}$.

Keep this "distribution over $\boldsymbol{p}$" picture in mind : we will call it Q and use it to break the decomposition

EDL: obtaining AU, EU and TU in closed form

The Dirichlet is the second-order distribution q, so we can plug it straight into the **same** information-theoretic decomposition. With $\boldsymbol{\alpha}$ the concentrations, $\alpha_0 = \sum_c \alpha_c$, ψ the digamma function, and mean prediction $\bar{\boldsymbol{p}} = \boldsymbol{\alpha}/\alpha_0$:

$$\mathrm{TU} = H[\bar{\boldsymbol{p}}] = -\sum_c \tfrac{\alpha_c}{\alpha_0} \ln \tfrac{\alpha_c}{\alpha_0} \qquad \text{entropy of the mean}$$

$$\mathrm{AU} = \mathbb{E}_{\boldsymbol{p}\sim\mathrm{Dir}(\boldsymbol{\alpha})}\big[H[\boldsymbol{p}]\big] = -\sum_c \tfrac{\alpha_c}{\alpha_0}\big(\psi(\alpha_c{+}1) - \psi(\alpha_0{+}1)\big) \qquad \text{expected entropy (closed form)}$$

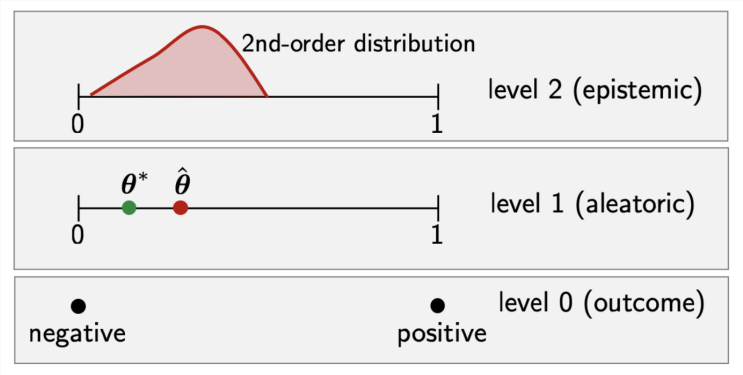
$$\mathrm{EU} = \mathrm{TU} - \mathrm{AU} \;=\; I(y;\boldsymbol{p}) \;\geq 0 \qquad \text{mutual information (residual)}$$

The original EDL paper (Sensoy et al.) used a simpler, evidence-based scheme:

- AU: the spread of the mean prediction, $H[\bar{\boldsymbol{p}}]$
- EU: the *evidence mass* $u = \frac{C}{\alpha_0}$ ($u \to 1$ with no evidence, $u \to 0$ as $\alpha_0 \to \infty$)
- TU: not explicitly decomposed.

Sensoy et al. 2018 Bengs et al. 2022 Mucsányi et al. 2024

Example in 2D from Wimmer et al.



With ϕ the probability of a positive prediction as in the Bernoulli model.

Wimmer et al., Quantifying Aleatoric and Epistemic Uncertainty in Machine Learning..., 2022

Recap: decompositions

Four ways to write TU = AU + EU

- **Information-theoretic:** $\mathbb{E}_q H[p]$ (avg. entropy) + $I(y; \boldsymbol{W} \mid x)$ (mutual info). *One* estimator gives both.
- **Variance-based:** law of total variance → $\mathbb{E}_q[\mathrm{Var}] + \mathrm{Var}_q[\mathbb{E}]$. Same derivation, but with variance. Natural for regression.
- **Kendall & Gal / two-head:** AU from a learned noise head, EU from $q(\boldsymbol{W})$.
- **Evidential (EDL):** a *distribution over distributions* from a single forward pass. AU = where it is centred, EU = how diffuse it is. No sampling, but weakly identified.

The first two are *decompositions* of a single quantity; the last two *construct* the two parts separately.

Recap: how each part is actually estimated

Aleatoric: spread at fixed W

- **Softmax entropy** directly: CE training $\Rightarrow$ the network learns $p(y \mid x)$.
- **Noise head**: predict $\sigma^2(x)$ alongside the mean (Kendall & Gal).
- **Generative models**: SSN, Prob. U-Net, PHiSeg, diffusion : sample *coherent* plausible outputs.
- **Evidential**: the mean of the predicted distribution over distributions.

Epistemic: spread over predictions

- **By sampling $W_m \sim q(W)$**: ensembles, MC dropout, SWAG, Laplace, BNN posterior. EU = *disagreement* between samples.
- **Directly**: EDL predicts the spread itself, no sampling.
- **Distance-aware**: distance to the training data in feature space (DUQ, SNGP).

Benchmarking the Decomposition

Does the decomposition work?

What *would* success look like? The framework suggests that:

The disentanglement promise

1. **AU** tracks data ambiguity → good for *ambiguity modelling* (AMB)
2. **EU** tracks distribution shift → good for *OOD detection* (OODD)
3. **TU** tracks overall risk → good for *calibration* (CAL)
4. The two components are **separable**: low correlation, each specialised to its task.

Does it actually work like that?

Mucsányi et al., NeurIPS 2024 Kahl et al. (ValUES), ICLR 2024

Several benchmarks, one consistent story

Over the last years, several papers have asked whether the decomposition actually works. They differ in domain, methods and metrics, but paint a similar picture.

Study	**Domain**	**Core question**
de Jong et al. 2024	Classification	Are AU and EU *orthogonal*?
Mucsányi et al. 2024	Classification (ImageNet)	Does any method give a *disentangled pair*?
Kahl et al. (ValUES) 2024	Segmentation	Does the *theoretically consistent* measure win?
Christensen et al. 2026 (**ours**)	Segmentation (medical)	What happens when we model AU *and* EU *together*?

Benchmark 1: de Jong et al. : orthogonality as the criterion

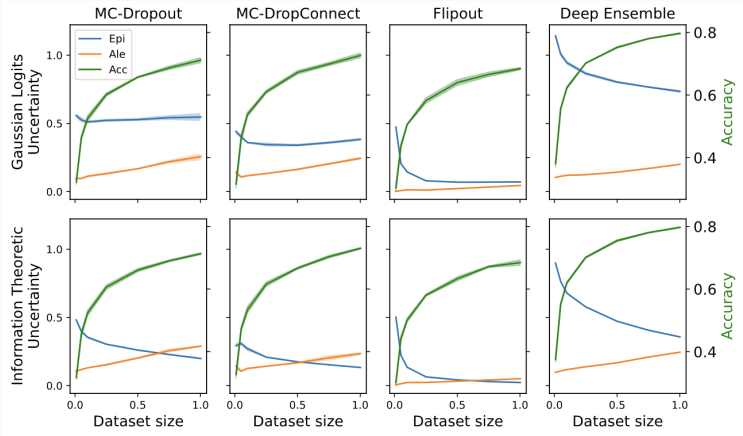
What they did.

- Argue that a correct decomposition needs **orthogonality**: each uncertainty must *not* be affected by the other. They argue that orthogonality together with consistency characterises disentanglement.
- Propose the **Uncertainty Disentanglement Error (UDE)** to measure it.
- Three controlled experiments where the *ground truth* is known by construction:
  - **Dataset size** $\downarrow$ $\Rightarrow$ only EU should rise
  - **OOD shift** $\Rightarrow$ only EU should rise
  - **Label noise** $\uparrow$ $\Rightarrow$ only AU should rise
- Compare two formulations: the *information-theoretic* split vs. the *Gaussian-logits* split, across MC dropout, DropConnect, deep ensembles, Flipout.

What they found. The information-theoretic approach disentangles *better*, but with both formulations each estimate remains **largely contaminated by the other**.

de Jong, Sburlea, Sabatelli, Valdenegro-Toro, arXiv:2408.12175

Benchmark 1: de Jong et al. : orthogonality as the criterion



(screenshot from paper)

de Jong, Sburlea, Sabatelli, Valdenegro-Toro, arXiv:2408.12175

Benchmark 2: Mucsányi et al. : the first disentanglement benchmark

What they did.

- Re-implemented a wide spectrum of uncertainty estimators in one codebase: *Bayesian* (ensembles, dropout, Laplace, SWAG), *evidential* (EDL, PostNet), and *deterministic* (Mahalanobis, DDU, DUQ). Everything on **ImageNet**.
- Evaluated them across a range of tasks: abstained / correctness prediction, OOD detection, and aleatoric estimation (predicting human label ambiguity).
- Measured the **rank correlation** between the AU and EU for each method.

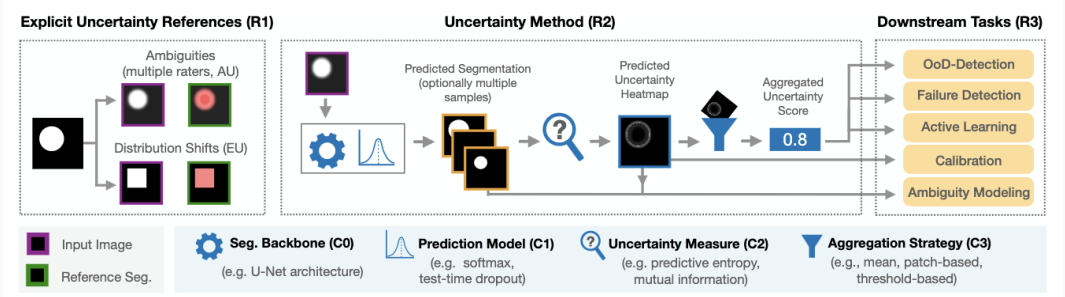
What they found.

- For **all** approaches the AU and EU components of a decomposition are highly rank-correlated : they capture *the same notion of uncertainty* in practice.
- **The way forward:** combine *separate* estimators, each strongly reflecting one source, rather than decomposing one predictive distribution.

Benchmark 3: Kahl et al. : ValUES

What they did.

- A framework for systematic validation of uncertainty estimation in semantic **segmentation**.
- Evaluates the UQ method plus configuration options, for different downstream tasks.
- Methods: softmax, deep ensembles, MC dropout, SSN, TTA.
- Evaluated on simulated data, Cityscapes / GTA V, and LIDC.



Benchmark 3: Kahl et al. : ValUES

What they found.

- On *simulated* data, methods recover the intended uncertainties well. On *real* data, AU and EU are entangled.
- The **theoretically consistent** measure is frequently *not* the best-performing one.
- The choice of *aggregation strategy* can matter as much as the choice of method.

(a)

Question	Qual./Quant.	Toy Dataset	LIDC	GTA5/CS
Q1: Do AU-measures capture AU?	Quantitative	yes	partially	partially
	Qualitative	yes	often (i.i.d.)	partially
Q2: Do EU-measures capture AU?	Quantitative	no	often yes	no
	Qualitative	no	often yes	no
Q3: Do EU-measures capture EU?	Quantitative	yes	yes	yes
Q4: Do AU-measures capture EU?	Quantitative	depends on AU in data	partially	mostly not

Kahl, Lüth, Zenk, Maier-Hein, Jaeger, ICLR 2024

Benchmark 3: Kahl et al. : limitations of ValUES

Methodological

- **Single source at a time.** ValUES uses AU (SSN/TTA) *or* EU (dropout/ensemble) methods in *isolation*, never combined, so the AU×EU *interplay* is never tested.
- **Decomposition without EU.** ValUES feeds aleatoric (latent-induced) samples into the *epistemic* slot of the **information-theoretic decomposition**.

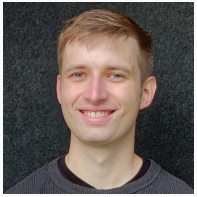
Data / experimental

- **GTA/Cityscapes label-flipping.** AU is simulated by random label flips (a given class always has the same AU). No dependence on image structure, minimal realism.
- **LIDC AU is degenerate.** Annotator disagreement is dominated by nodule *presence/absence*, not boundary delineation.
- **OOD that is also aleatoric.** The OOD split is based on texture and malignancy, which selects images that are also ambiguous (e.g. blurry) and therefore carry AU too.

Benchmark 4: Christensen et al. (ours) : motivation

We address the limitations of prior work with four contributions:

- **Model both sources, together.** We evaluate the full **AU**×**EU** grid (4×5 combinations), where ValUES used AU *or* EU methods in isolation. This lets us study their *interplay* and whether modelling both is advantageous.
- **An entanglement metric Δ.** We quantify whether the theoretically *correct* measure actually outperforms the *wrong* one for a task.
- **Real, not simulated, ambiguity.** We use *multi-annotator* medical data (LIDC, Chákṣu), where AU comes from rater disagreement, rather than synthetic data.
- **Diagnosis and recommendations.** We analyse the *sources* of entanglement and give task-specific model recommendations balancing entanglement and performance.



Jakob Lønborg Christensen (PhD student at DTU). Work done during research visit in Lucerne.

Benchmark 4: Christensen et al. (ours) : setup

AU methods (4)

Softmax, SSN, Prob. U-Net, diffusion

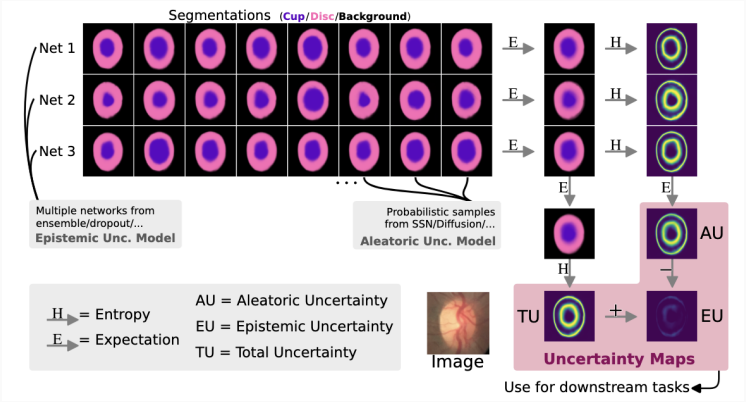
EU methods (5)

No EU, MC dropout, SWAG-D, SWAG, deep ensemble

Protocol

- **19** of the $4 \times 5 = 20$ combinations are valid $\times$ **3** datasets $\times$ **3** tasks $\times$ **5** seeds
- Shared attention-U-Net backbone ($\approx$16 M params) to remove architecture effects
- 10 AU samples $\times$ 10 EU instances = up to 100 predictions per image

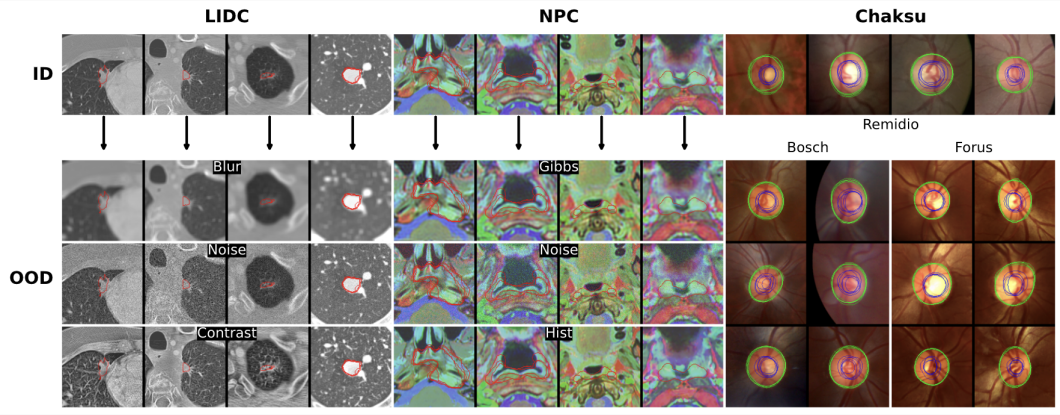
Benchmark 4: Christensen et al. (ours) : modelling both sources (reminder)



$$\mathrm{AU} = \frac{1}{M}\sum_m H[\bar{\pi}_m], \qquad \mathrm{TU} = H\Big[\frac{1}{M}\sum_m \bar{\pi}_m\Big], \qquad \mathrm{EU} = \mathrm{TU} - \mathrm{AU}.$$

Christensen et al. arxiv (2026)

Benchmark 4: Christensen et al. (ours) : data



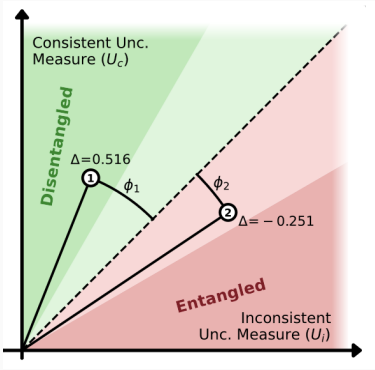
Data (all multi-annotator, AU has a GT proxy), plus synthetic and device-shift OOD sets.
Tasks: OODD (AUROC) → EU; AMB (NCC vs. rater variance) → AU; CAL (ECE) → TU.

Benchmark 4: Christensen et al. (ours) : evaluation metric

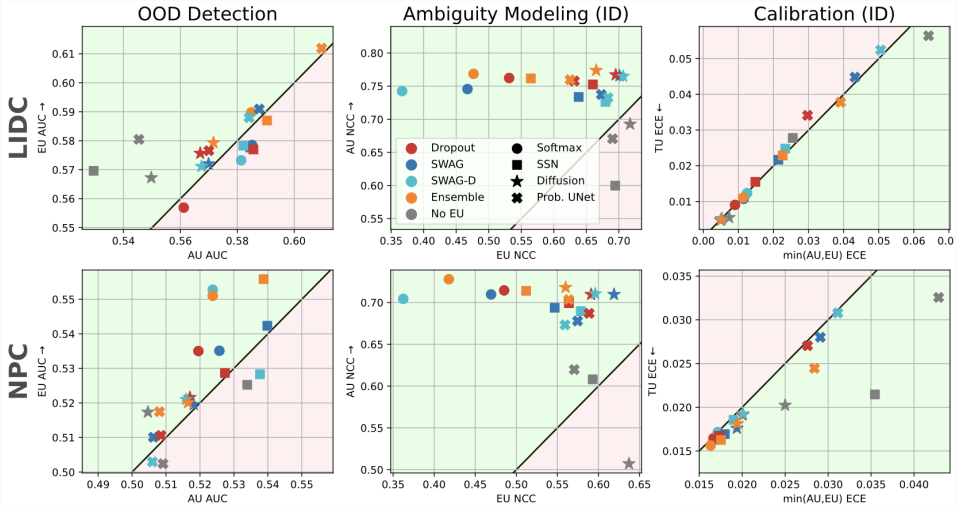
If the decomposition worked, the *theoretically consistent* measure U_c should beat the *inconsistent* one U_i on each task. We turn that comparison into a number:

$$\Delta \;=\; s\,\frac{\arctan(U_c/U_i)-\pi/4}{\pi/4} \;=\; s\,\frac{\phi}{\pi/4}, \qquad s=\begin{cases}-1 & \text{lower is better (ECE)}\\ +1 & \text{otherwise}\end{cases}$$

- For $U_c, U_i > 0$: $\Delta \in (-1, 1)$; it measures the signed *angle* ϕ from the $U_c = U_i$ diagonal.
- $\Delta > 0$: consistent measure wins $\Rightarrow$ **disentangled**.
- $\Delta < 0$: inconsistent measure wins $\Rightarrow$ **entangled**.

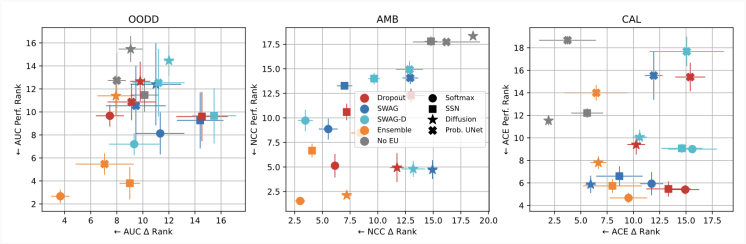


Benchmark 4: Christensen et al. (ours) : results



Benchmark 4: Christensen et al. (ours) : results

1. **The theory holds weakly.** U_c beats U_i for the majority of models and tasks, but many combinations sit close to the diagonal, i.e. barely disentangled.
2. **The combination does not behave as theory predicts.** If AU and EU were disentangled, OODD should be driven by the EU method and AMB by the AU method. Instead, points cluster by AU *method type* even for OODD.
3. **Epistemic collapse is pervasive.** Mean EU/AU < 0.2 for every combination with an EU component $\Rightarrow$ TU $\approx$ AU, which entangles TU with AU by construction. Worst for dropout and SWAG-D, least bad for ensembles.



Benchmark 4: Christensen et al. (ours) : results

4. **Ensembles work well.** Best performance *and* lowest entanglement. Likely because independent training trajectories give more diverse predictions.
5. **A plain softmax ensemble is the winner** overall, despite its conceptual simplicity.

	Performance	Entanglement (Δ)	Both
Out-of-distribution detection (OODD)	SSN ens. /Softmax ens.	Softmax ens.	Softmax ens.
Amb. Mod. (AMB)	Diffusion ens. /Softmax ens.	SSN ens. /Softmax ens.	Softmax ens.
Calibration (CAL)	Softmax ens.	Diffusion SWAG	Diffusion SWAG /Softmax ens.
Overall	Softmax ens.	Softmax ens.	Softmax ens.

6. **Do not decompose without an EU component.** "No EU" models are consistent outliers; applying the information-theoretic decomposition to them (as ValUES does) lacks a theoretical foundation.

Benchmarking recap

The four benchmarks paint a consistent picture

The aleatoric/epistemic decomposition does *not* reliably deliver two separable, task-specialised measures. Across different domains, different tasks, and different benchmarking protocols, all studies show that AU and EU are strongly correlated in practice.

Why does it not work?

Why Does the Theory Fail?

Overview

Going from mild to more fundamental objections:

0. **Conceptual confusion.** Additivity is not orthogonality : the identity never promised disentanglement.
1. **Estimation failure.** Epistemic collapse: the $q(\boldsymbol{W})$ may end of very *concentrated*.
2. **Wrong measures.** The entropy of the prediction doesn't care about the spread of Q; mutual information measures *conflict*, not *ignorance*.
3. **Over-constrained Structure.** No pair of "ideal" measures on a common scale can satisfy $\mathrm{TU} = \mathrm{AU} + \mathrm{EU}$.
4. **Unclear definitions.** It may not always be clearly defined what even is epistemic and aleatoric in certain cases.

Problem 0: additivity is not orthogonality

It is tempting to interpret $\mathrm{TU} = \mathrm{AU} + \mathrm{EU}$ this as "the two are separated." But the equation doesn't say that.

Misconception

Additivity says the parts *sum* to the whole. It says **nothing** about their *correlation* across inputs.

- AU and EU can be highly correlated without violating additivity.
- In the simplest case, both AU and EU are functionals of $\{\pi_m\}$, so not entirely surprising that they are correlated.
- $\mathrm{EU} = \mathrm{TU} - \mathrm{AU}$ is defined as residual, so the two terms are algebraically directly connected.

Problem 1: epistemic collapse

In modern *overparametrised* networks the approximate posterior $q(\boldsymbol{W})$ can often concentrate, i.e. ensemble members / dropout masks give *nearly identical* predictive functions on in-distribution data.

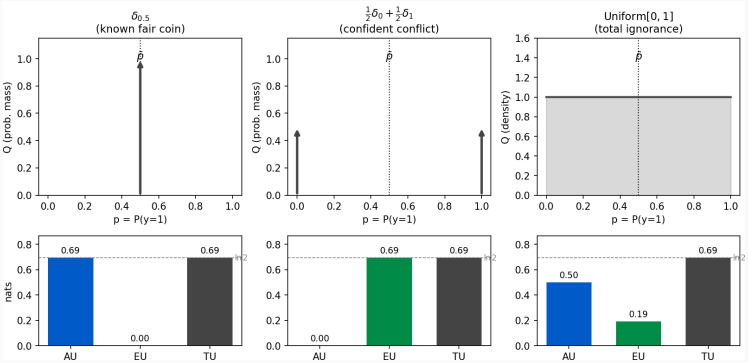
$$\mathrm{Var}_{q(\boldsymbol{W})}\big(\mathbb{E}[y \mid \boldsymbol{W}, x]\big) \approx 0 \implies \text{EU} \approx 0 \implies \text{TU} \approx \text{AU}.$$

- Since EU $=$ TU $-$ AU, TU becomes mechanically correlated with AU (recall in our study EU/AU < 0.2 across *every* combination).
- Fellaji & Pennerath (2024) show that it worsens with model size (but currently not yet entirely clear why).

Kirsch, TMLR 2025 Fellaji & Pennerath, 2024

A toy example for the next points

Binary problem, $p = P(y = 1)$. Write q for the *second-order* distribution over p (like before for the EDL slide). Note the distribution here is over the predictions, not over $\mathbf{W}$). Three cases:



Calculation of TU, AU and EU

$$\underbrace{H\big[\mathbb{E}_q\, p\big]}_{\text{TU}} = \underbrace{\mathbb{E}_q\, H[\,p\,]}_{\text{Aleatoric (AU)}} + \underbrace{I(y^\star; \boldsymbol{W} \mid x^\star)}_{\text{Epistemic (EU)}}$$

q	$\overline{p}$	$TU = H[\mathbb{E}_q p]$	$\text{AU} = \mathbb{E}_q[H(p)]$	$\text{EU} = TU - AU$
$\delta_{0.5}$ (known fair coin)	0.5	0.69	0.69	0
$\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ (confident conflict)	0.5	0.69	0	**0.69**
$\mathrm{Uniform}[0,1]$ (total ignorance)	0.5	0.69	0.5	0.19

Two things to notice, which we unpack on the next two slides:

- TU $= H[\,\overline{p}\,] = -0.5 \log 0.5 - 0.5 \log 0.5 = \log 2 = 0.69$ for *all three* rows.
- EU is *largest* for confident conflict, not for genuine ignorance.

Adapted from: Wimmer, Sale, Hofman, Bischl, Hüllermeier, UAI 2023

Problem 2a: total uncertainty sees only the mean of Q

$$\mathrm{TU} = H\big[\,\mathbb{E}_q[\,p\,]\,\big] \quad \text{depends on } Q \textit{ only} \text{ through its expectation } \overline{p}.$$

This means **any** mean-preserving spread of q leaves TU unchanged:

$$\delta_{0.5} \;\; \text{(perfect knowledge)} \qquad \text{and} \qquad \mathrm{Uniform}[0,1] \;\; \text{(total ignorance)}$$

have the *same* $\overline{p} = 0.5 \Rightarrow$ the same $\mathrm{TU} = \log 2$.

Violation of invariance to mean-preserving spread

Entropy of the predictive violates a basic "mean-preserving spread" requirement as postulated by Wimmer et al. It is blind to the concentration of q. The concentration of q is exactly what should tell us the epistemic uncertainty, so there is something fundamentally wrong with measuring uncertainty like this.

Wimmer et al., UAI 2023

Problem 2b: mutual information is conflict, not ignorance

From the toy table: MI ranks *confident conflict* (0.69) above *total ignorance* (0.19).

Mutual information measures the wrong thing

$MI(y; \boldsymbol{W} \mid x)$ is larger when hypotheses *disagree confidently* than for *total ignorance* (uniform). This is expected behaviour for the MI (in the two delta example with gain the most information by observing one of the delta peaks). However, this is not what we expect from a measure of uncertainty. Total ignorance should have maximum EU!

MI is not answering the question "how little do I know?". It is fine measure to measure *expected information gain* (as in active learning). But, it should not be used to interpret *how much we do not know*.

Wimmer et al., UAI 2023

Problem 3: the additive structure is over-constrained

Because AU $=$ TU $-$ EU, AU is never measured directly. It is whatever EU leaves behind, so it inherits EU's errors. **Example: Start of training**. With no data, EU is (correctly) maximal and, on a shared scale, EU $\approx$ TU. Additivity then forces

$$\text{AU} = \text{TU} - \text{EU} \approx 0,$$

Uncertainty has to be positive, so this split forces AU to 0.

TU cannot decompose additively into EU and AU if AU is positive.

For "ideal" measures on a shared scale, TU $=$ AU $+$ EU cannot hold exactly. The decomposition recovers only TU $=$ EU $+$ (a lower bound on AU). This still holds at other training stages, not just at initialisation.

Wimmer et al., UAI 2023

Problem 4: Unclear definitions

The aleatoric/epistemic split sounds clean, but its definitions *contradict each other*, even in toy examples.

Epistemic: disagreement or number of models?

Belief has collapsed to two models, $\theta = 0$ or $\theta = 1$ (two Diracs). Is EU high or low?

- *Disagreement view* (MI; Houlsby): they maximally disagree $\Rightarrow$ EU **maximal**.
- *Number-of-models view* (Wimmer): only two models remain $\Rightarrow$ EU **near-minimal**

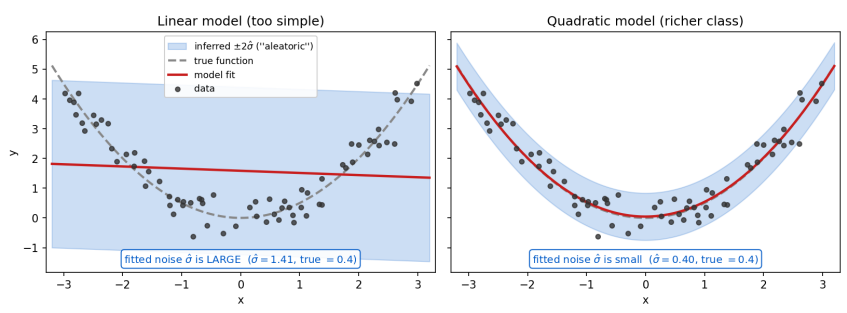
Aleatoric: can model class be used to reduce it?

Fit a linear model to quadratic data. Is the leftover "model bias" aleatoric?

- *Bayes-optimal-within-class view*: irreducible even with infinite data $\Rightarrow$ aleatoric.
- *Data-uncertainty view*: a richer model class removes it $\Rightarrow$ epistemic.

Kirchhof, Kasneci, Kasneci, ICML 2025 (Position); Wimmer et al. 2023; Hüllermeier & Waegeman 2021

Problem 4: Unclear definitions



Problem 4: Unclear definitions

Even granting definitions, the two cannot be cleanly separated, and under *interaction*.

Under interaction things become weird

An LLM agent can ask follow-up questions. Then:

- Apparent epistemic (missing knowledge) becomes aleatoric if questions *fail* to reduce it.
- Apparent aleatoric (an underspecified query) becomes epistemic, reducible *by asking*.

Der Kiureghian & Ditlevsen: calling an uncertainty aleatoric or epistemic is ultimately a *subjective modeller's choice* of what to reduce, not a true distinction.

Kirchhof, Kasneci, Kasneci, ICML 2025; Mucsányi et al. 2024; Gruber et al. 2023; Der Kiureghian & Ditlevsen 2009

Does using separate sources fix this?

A natural hope is that if AU and EU come from *different mechanisms*, the problems should not apply. This is what Kendall & Gal do, and what we did in the Christensen et al. benchmark paper.

Separate AU/EU modeling doesn't fix the decomposition's problems

The two sources are still *combined through the information-theoretic decomposition*. Even though the estimators change, the uncertainties are still combined using the flawed additive decomposition.

de Jong et al. confirm this empirically: even the heteroscedastic (Gaussian-logits) split stays entangled. **Possible way out:** treat AU and EU as *unrelated quantities*, each validated on its own task. E.g. EU as Mahalanobis distance in feature space, AU as sample diversity from a diffusion segmentation model. But then no way to estimate TU.

Kendall & Gal, NeurIPS 2017; de Jong et al., arXiv:2408.12175; Christensen et al. arxiv (2026)

What is valid and where do the problems start?

Not everything about AU and EU is broken. The concept is still practically important.

- The original decomposition for the prediction probability is still valid.

$$p(y^\star \mid x^\star, \mathcal{D}) \;=\; \int p(y^\star \mid \boldsymbol{W}, x^\star)\, q(\boldsymbol{W})\, d\boldsymbol{W} \;=\; \mathbb{E}_{q(\boldsymbol{W})}[\, p(y^\star \mid \boldsymbol{W}, x^\star)\,]\,.$$

- The trouble begins when we try to *split* it additively using entropy or variance. We try to write the decomposition as an addition when the terms may in reality interact non-linearly.

Possible avenues for fixing this: credals sets, proper scoring rules, ... However, nothing currently solves all problems and *just works*.

So why did I teach you all of this?

The distinction between AU and EU still matters!

Knowing *what kind* of uncertainty you face changes what you should *do*. Irreducible ambiguity in the image means deferring to a human or accepting a set-valued answer. Ignorance about an unfamiliar input means collecting data, retraining, or refusing to predict.

- People have been working on UQ for neural networks for decades and still there are no tools that work reliably.
- But, the problem matters enormously for clinical deployment.
- Everything in this lecture is the foundation you need to build something better.

Practical advice, if you just want to use uncertainty

- For a general UQ task, **prefer total uncertainty as your one number.** $H[p(y \mid x^\star)]$ is more robust in practice than AU and EU alone.
- **If you need AU and EU, model them separately.** Two estimators each for its own task (e.g. OODD, ambiguity detection) is often better. But, then you can't get TU.
- **Ensembles often work well.** A deep ensemble with softmax AU estimation worked best in our benchmark study and similar findings in many other studies.
- If you do model AU and EU, **calculating their correlation** is a simple baseline test.
- Because our uncertainty measures are flawed, **frameworks that give distribution-free guarantees** *regardless of how good the underlying score is* (e.g. like conformal prediction and risk controlling prediction sets) are a promising direction.

Mucsányi et al., NeurIPS 2024; Valdenegro-Toro & Mori, CVPRW 2022

References

References i

- Kendall & Gal. *What uncertainties do we need in Bayesian deep learning for computer vision?* NeurIPS 2017.
- Depeweg, Hernández-Lobato, Doshi-Velez, Udluft. *Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning.* ICML 2018.
- Hüllermeier & Waegeman. *Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods.* Machine Learning, 2021.
- Cover & Thomas. *Elements of information theory.* 2nd ed., Wiley, 2006.
- Houlsby, Huszár, Ghahramani, Lengyel. *Bayesian active learning for classification and preference learning.* arXiv:1112.5745, 2011.
- Gal, Islam, Ghahramani. *Deep Bayesian active learning with image data.* ICML 2017.
- MacKay. *A practical Bayesian framework for backpropagation networks.* Neural Computation, 1992.
- Neal. *Bayesian learning for neural networks.* Springer, 1996.

References ii

- Blundell, Cornebise, Kavukcuoglu, Wierstra. *Weight uncertainty in neural networks.* ICML 2015.
- Gal & Ghahramani. *Dropout as a Bayesian approximation.* ICML 2016.
- Osband. *Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout.* NIPS Workshop on Bayesian Deep Learning, 2016.
- Folgoc et al. *Is MC dropout Bayesian?* arXiv:2110.04286, 2021.
- Wan, Zeiler, Zhang, LeCun, Fergus. *Regularization of neural networks using DropConnect.* ICML 2013.
- Wen, Vicol, Ba, Tran, Grosse. *Flipout: efficient pseudo-independent weight perturbations on mini-batches.* ICLR 2018.
- Lakshminarayanan, Pritzel, Blundell. *Simple and scalable predictive uncertainty estimation using deep ensembles.* NeurIPS 2017.
- Izmailov, Podoprikhin, Garipov, Vetrov, Wilson. *Averaging weights leads to wider optima and better generalization.* UAI 2018. (SWA)

References iii

- Maddox, Garipov, Izmailov, Vetrov, Wilson. *A simple baseline for Bayesian uncertainty in deep learning.* NeurIPS 2019.
- Martens & Grosse. *Optimizing neural networks with Kronecker-factored approximate curvature.* ICML 2015. (KFAC)
- Daxberger, Kristiadi, Immer, Eschenhagen, Bauer, Hennig. *Laplace Redux: effortless Bayesian deep learning.* NeurIPS 2021.
- Lee, Lee, Lee, Shin. *A simple unified framework for detecting out-of-distribution samples and adversarial attacks.* NeurIPS 2018.
- Van Amersfoort, Smith, Teh, Gal. *Uncertainty estimation using a single deep deterministic neural network.* ICML 2020.
- Liu, Lin, Padhy, Tran, Bedrax-Weiss, Lakshminarayanan. *Simple and principled uncertainty estimation with deterministic deep learning via distance awareness.* NeurIPS 2020.

References iv

- Mukhoti, Kirsch, van Amersfoort, Torr, Gal. *Deep deterministic uncertainty: a new simple baseline.* CVPR 2023.
- Nix & Weigend. *Estimating the mean and variance of the target probability distribution.* ICNN 1994.
- Sohn, Lee, Yan. *Learning structured output representation using deep conditional generative models.* NeurIPS 2015.
- Oktay et al. *Attention U-Net: learning where to look for the pancreas.* MIDL 2018.
- Kohl et al. *A probabilistic U-Net for segmentation of ambiguous images.* NeurIPS 2018.
- Kohl et al. *A hierarchical probabilistic U-Net for modeling multi-scale ambiguities.* arXiv:1905.13077, 2019.
- Baumgartner et al. *PHiSeg: capturing uncertainty in medical image segmentation.* MICCAI 2019.
- Monteiro et al. *Stochastic segmentation networks: modelling spatially correlated aleatoric uncertainty.* NeurIPS 2020.

References v

- Ayhan & Berens. *Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks.* MIDL 2018.
- Amit, Shaharbany, Nachmani, Wolf. *SegDiff: image segmentation with diffusion probabilistic models.* arXiv:2112.00390, 2021.
- Wolleb, Sandkühler, Bieder, Valmaggia, Cattin. *Diffusion models for implicit image segmentation ensembles.* MIDL 2022.
- Rahman, Valanarasu, Hacihaliloglu, Patel. *Ambiguous medical image segmentation using diffusion models.* CVPR 2023.
- Sensoy, Kaplan, Kandemir. *Evidential deep learning to quantify classification uncertainty.* NeurIPS 2018.
- Amini, Schwarting, Soleimany, Rus. *Deep evidential regression.* NeurIPS 2020.
- Bengs, Hüllermeier, Waegeman. *Pitfalls of epistemic uncertainty quantification through loss minimisation.* NeurIPS 2022.

References vi

- Charpentier, Zügner, Günnemann. *Posterior network: uncertainty estimation without OOD samples via density-based pseudo-counts.* NeurIPS 2020. (PostNet)
- Christensen et al., *Diffusion Based Ambiguous Image Segmentation.* Scandinavian Conference on Image Analysis 2025.
- Christensen et al. 2026 arxiv. *Rethinking Uncertainty Quantification and Entanglement in Image Segmentation.* arXiv, 2026.
- Guo, Pleiss, Sun, Weinberger. *On calibration of modern neural networks.* ICML 2017. (ECE)
- de Jong, Sburlea, Sabatelli, Valdenegro-Toro. *How disentangled are your classification uncertainties?* arXiv:2408.12175, 2024.
- Mucsányi, Kirchhof, Oh. *Benchmarking uncertainty disentanglement: specialized uncertainties for specialized tasks.* NeurIPS 2024 (Datasets & Benchmarks).
- Kahl, Lüth, Zenk, Maier-Hein, Jaeger. *ValUES: a framework for systematic validation of uncertainty estimation in semantic segmentation.* ICLR 2024.

References vii

- Wimmer, Sale, Hofman, Bischl, Hüllermeier. *Quantifying aleatoric and epistemic uncertainty: are conditional entropy and mutual information appropriate measures?* UAI 2023.
- Kirsch. *(Implicit) ensembles of ensembles: epistemic uncertainty collapse in large models.* TMLR 2025.
- Fellaji & Pennerath. *The epistemic uncertainty hole: an issue of Bayesian neural networks.* 2024.
- Gruber, Schenk, Schierholz, Kreuter, Kauermann. *Sources of uncertainty in machine learning: a statisticians' view.* arXiv:2305.16703, 2023.
- Der Kiureghian & Ditlevsen. *Aleatory or epistemic? Does it matter?* Structural Safety, 31(2):105–112, 2009.
- Valdenegro-Toro & Saromo Mori. *A deeper look into aleatoric and epistemic uncertainty disentanglement.* CVPR Workshops (CVPRW), 2022.
- Kuhn, Gal, Farquhar. *Semantic uncertainty: linguistic invariances for uncertainty estimation in natural language generation.* ICLR 2023.

References viii

- Kirchhof, Kasneci, Kasneci. *Position: uncertainty quantification needs reassessment for large-language model agents.* ICML 2025.
- Zepf et al. *Laplacian segmentation networks improve epistemic uncertainty quantification.* MICCAI 2024.

Datasets

- Armato et al. *The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans.* Medical Physics, 2011.
- Kumar et al. *Chákṣu: a glaucoma specific fundus image database.* Scientific Data, 2023.
- Cordts et al. *The Cityscapes dataset for semantic urban scene understanding.* CVPR 2016.
- Richter, Vineet, Roth, Koltun. *Playing for data: ground truth from computer games.* ECCV 2016. (GTA V)