Trust Me, I Know the Way: Predictive Uncertainty in the Presence of Shortcut Learning

Lisa Wimmer, Bernd Bischl & Ludwig Bothmann
Department of Statistics, LMU Munich; Munich Center for Machine Learning (MCML)
{firstname.lastname}@stat.uni-muenchen.de

Abstract

The correct way to quantify predictive uncertainty in neural networks remains a topic of active discussion. In particular, it is unclear whether the state-of-the art entropy decomposition leads to a meaningful representation of model, or *epistemic*, uncertainty (EU) in the light of a debate that pits *ignorance* against *disagreement* perspectives. We aim to reconcile the conflicting viewpoints by arguing that both are valid but arise from different learning situations. Notably, we show that the presence of *shortcuts* is decisive for EU manifesting as disagreement.

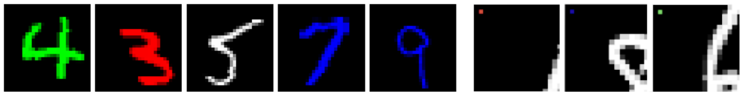

Figure 1: Examples of 3-class CMNIST3 data (*left*) with full-image color shortcut and PMNIST3 (*right*; cropped for better visibility) with 1×1 colored patches (shortcuts in 95% of images).

1 Introduction

Safety-critical applications require models to report their uncertainty about predicted outcomes. The rise of deep learning, with its powerful-but-overconfident algorithms, has sparked rapid progress in uncertainty quantification (UQ, Guo et al., 2017; Gawlikowski et al., 2023). Yet, the problem of faithfully representing, measuring and communicating uncertainty remains far from solved (Van Der Bles et al., 2019; Bickford Smith et al., 2024). One disputed aspect is the correct approach to disaggregate uncertainty into its *aleatoric (AU)* and *epistemic (EU)* components (Hüllermeier and Waegeman, 2021; Gruber et al., 2023). AU arises when the features X do not suffice to predict the target beyond doubt—e.g., due to omitted variables—even if we had access to infinite data. The ground-truth distribution over labels Y, $\boldsymbol{\theta}^* = p(Y|X = \boldsymbol{x})$, will thus generally have positive dispersion (i.e., AU). EU, on the other hand, surrounds the approximation quality of the model $h(\boldsymbol{x}) = \hat{p}_n(Y|X = \boldsymbol{x})$ based on n observations. EU is reducible in the sense that $\hat{p}_\infty$ recovers p[1].

Attributing predictive uncertainty to its sources helps understand learning dynamics and provides an inroad for targeted model improvement. The widely-used decomposition of Shannon entropy[2] (Shannon, 1948) as a measure of total uncertainty (TU) offers a neat mathematical expression but has attracted criticism for conflating distinct concepts (Wimmer et al., 2023; Schweighofer et al., 2024; Bickford Smith et al., 2024) and being ineffective in practice (Mucsányi et al., 2024; Fellaji and Pennerath, 2024). It is unclear, in particular, how well the resulting EU measure reflects lack of knowledge. Besides the uncertainty of a single model $h \in \mathcal{H}$ about the predicted outcome (AU), there is uncertainty about which model even is the correct one (EU). Quantifying EU thus demands a bi-level framework that accounts for different hypotheses, uncertainty rising with the number of hypotheses deemed plausible (Hofman et al., 2024). The EU representation problem concerns this multiplicity and admits two competing viewpoints: Maximum uncertainty is attained when (V1) all hypotheses are equally likely *or* (V2) only hypotheses reflecting full confidence for one label each

[1] We often assume a correctly specified model, equating hypotheses with parameterizations (Draper, 1995).

[2] Entropy is typically used with categorical target distributions. An analogous decomposition exists for variance-based UQ in regression tasks (Depeweg et al., 2018; Sale et al., 2024).

are assigned nonzero probability (in complete disagreement). Take the toy example of predicting a coin toss with EU concerning coin bias. Should EU be highest when the model deems any degree of bias equiprobable, or when it is sure the coin shows either heads or tails always but not which? Clarity about this worst-case scenario is necessary in uncertainty-based tasks like active learning (Smith et al., 2023) and out-of-distribution (OOD) detection (Azizmalayeri et al., 2024). V1 fits a classical Bayesian view of associating uninformed beliefs with uniformity (Dubois et al., 1996), while EU as per the entropy decomposition embodies V2, becoming maximal when all hypotheses express full confidence in utter disagreement (Shoja and Soofi, 2017; Wimmer et al., 2023).

We do not aspire to resolve the philosophical dilemma but argue instead that the two EU manifestations arise from different scenarios. More precisely, we postulate that conflicting hypotheses (V2) only emerge in the presence of *shortcut learning (SCL)*: One hypothesis at most recovers the true mapping while the others necessarily point to spurious patterns. Our experiments suggest that shortcuts in the data indeed prompt such disagreement. This observation bears important insights for learning dynamics and has not, to the best of our knowledge, been studied despite ample discussion about robust generalization (Nagarajan et al., 2021; Wald et al., 2021; Richens and Everitt, 2024).

2 Background & Related Work

Uncertainty Quantification Bayesian methods come with a natural bi-level uncertainty representation, posing a second-order probability distribution Q over first-order distributions $\boldsymbol{\theta}$ induced by hypotheses h (for details, see App. A.1), and have therefore risen to gold standard in UQ (Izmailov et al., 2021). We consider the special case of finite ensembles, which can be viewed as approximately Bayesian (Wilson and Izmailov, 2020; Wild et al., 2023; Mlodozeniec et al., 2024), but our argumentation holds for any approach expressing EU via multiple predictions (induced by a distribution or set of hypotheses; Hofman et al., 2024) per instance. Following the entropy ($H(\cdot)$) decomposition popularized by Houlsby et al. (2011) and Kendall and Gal (2017), EU in the Bayesian setting is measured via *mutual information*:

$$\text{EU} = \underbrace{I(Y;\Theta)}_{\text{mutual information}} = \underbrace{H(\mathbb{E}_Q[Y|\Theta])}_{\text{entropy (TU)}} - \underbrace{\mathbb{E}_Q[H(Y|\Theta)]}_{\text{conditional entropy (AU)}} = \mathbb{E}_Q\left[D_{\mathrm{KL}}\left(p(Y|\boldsymbol{\theta}) \,\|\, \bar{\boldsymbol{\theta}}\right)\right], \quad (1)$$

where Θ denotes the random variable of first-order probability distributions. Mutual information is equivalent to *Jensen-Shannon divergence* in the finite-ensemble case. The emphasis on disagreement is obvious from the Kullback-Leibler divergence term: $D_{\mathrm{KL}}(\cdot)$ increases with deviation between individual hypotheses $p(Y|\boldsymbol{\theta})$ and the consensus prediction $\bar{\boldsymbol{\theta}}$ (Eq. 3, 5-6; Shoja and Soofi, 2017).

Shortcut Learning SCL is fundamentally a problem of distribution shift and occurs when patterns picked up in training do not carry over to OOD scenarios (see Steinmann et al., 2024, for a comprehensive discussion). Relevant works discern *stable* and *unstable* features $X_s, X_u \subseteq X$, with X_u non-causally correlated to Y, in varying terminology (e.g., Chalupka et al., 2015; Eastwood et al., 2023). Those distinctions imply the desirability of predicting Y from X_s[3], which is not always possible[4]. Relying on X_u instead induces *shortcuts* that work during training but break at deployment in OOD environments. Some shortcuts mirror real-world spurious correlations[5], others are introduced during data collection. The patterns can be subtle—e.g., high-frequency noise invisible to the human eye—and easily go undetected. Shortcuts abound across data modalities and affect downstream concerns like adversarial robustness and fairness (Geirhos et al., 2020).

Bridging the Gap Now, what makes models succumb to shortcuts? Steinmann et al. (2024) list ill-defined tasks and noisy X_s as potential causes. More importantly, the same inductive biases we praise for enabling (in-distribution) generalization also encourage SCL. Neural networks (NNs) are especially susceptible: In extracting latent features that a fully-connected module can digest, it is only rational under Occam's principle to rely on shortcuts when those induce the simplest risk-minimizing data representation[6] (Geirhos et al., 2020; Friedrich et al., 2023). Some authors

[3] Unstable features, while poor predictors on their own, can still boost performance (Eastwood et al., 2023).

[4] Jalaldoust et al. (2024) argue that situations of *non-transportability*, when X_s is not available and more (of the same) data cannot alleviate the shortcut problem, amount to a form of irreducible AU.

[5] Associating cows with grass is reasonable but useless if failing to identify cows in other surroundings.

[6] Strong information compression from input to latent space can signal shortcuts (Adnan et al., 2022).

argue even that SCL cannot safely be avoided without a causal framework (Schölkopf et al., 2021). Considerable effort has been made in this spirit by moving beyond standard supervised settings. Notably, a strand of recent work exploits the inherent multi-basin dynamics of ensembles with additional diversity-boosting components (e.g., special regularizers (Teney et al., 2022) or exposure to OOD training data (Pagliardini et al., 2023; Scimeca et al., 2024)). In that sense, we connect the dots between a principled discussion about disagreement to represent EU and algorithms using some pragmatic notion of uncertainty as a practical tool.

3 Predictive Uncertainty in the Presence of Shortcut Learning

In short, we observe that uncertainty estimates based on data with explicit shortcuts differ from those without. We consider classification tasks derived from MNIST (LeCun et al., 1998, pooling digits 1-3, 4-6, 7-9) and solved using deep ensembles of size 3 (Lakshminarayanan et al., 2017, see App. A.2 for details). Shortcuts of strength s are introduced in $s\%$ of samples by coloring entire images (CMNIST3) or adding 1×1 colored pixels (PMNIST3) according to class[7] (Fig. 1).

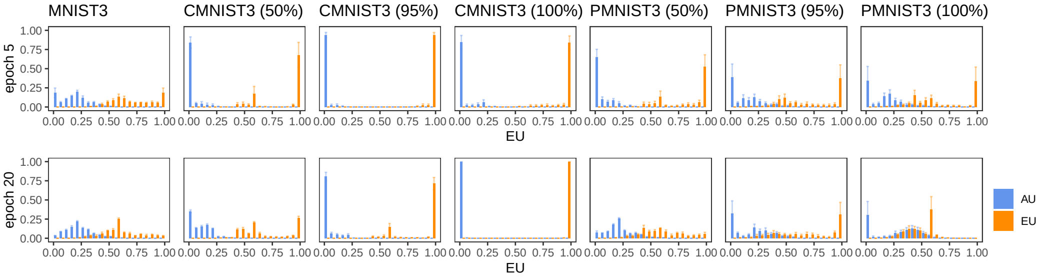

Figure 2: Uncertainty estimates (20% least-confident predictions) based on data with varying s; ensemble size 3; average over 3 independent runs (error bars: ± 1 standard error).

Fig. 2 shows the distribution of uncertainty estimates for MNIST3, CMNIST3 and PMNIST3 ($s \in \{50, 95, 100\}$). The respective test data consist of original class-0 images not seen during training. Uncertainty under SCL manifests as near-maximal mutual information[8] (orange bars) for CMNIST3, at near-full confidence (meaning zero AU; blue bars), under strong-enough shortcuts (columns 3-4). In other words, the ensemble members produce conflicting predictions with high probability on different classes for most test images (less so when the shortcut is only partial). This behavior follows the disagreement interpretation of EU (V2 in Sec. 1). There is a similar, if weaker, pattern for the less-pronounced PMNIST3 cue. Hypotheses start to agree again later in a sort of model collapse for PMNIST3 with $s = 100$ (note that estimates evolve as training progresses; see also App. A.2.2). The shortcut-free MNIST3 task produces a diffuse distribution in both uncertainty components, reflecting varying degrees of confidence and disagreement across different test images (V1). Only few samples provoke full disagreement. Recall, however, that the test images are OOD for all datasets. Using EU estimates as a signal in OOD detection would then yield quite different interpretations depending on whether SCL is at play. Taking uncertainty estimates at face value can thus be deceptive without a clear understanding of what EU via disagreement implies.

4 Conclusion

Much of recent research in UQ concerns methodological improvement. While indisputably important, this narrow focus—when it entails silent acceptance of assumptions like i.i.d. data—isolates the field from more fundamental discussions. The confusion about the appropriate representation of EU arises from such a gap. We reconcile different views on representing EU (conflict *vs* ignorance) by showing how shortcuts, which would not exist if the i.i.d. assumption actually held, affect uncertainty estimates in support of the conflict-type notion. With this, we hope to contribute to a comprehensive view on generalization that stands the test of real-world learning situations.

[7] Our code will be made available upon publication.

[8] We normalize all uncertainty components to values in [0, 1] according to Eq. 2.

Acknowledgments

This paper is supported by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the German Federal Ministry of Education and Research.

References

M. Adnan, Y. Ioannou, C.-Y. Tsai, A. Galloway, H. R. Tizhoosh, and G. W. Taylor. Monitoring Shortcut Learning using Mutual Information. In *ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability*, 2022.

C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning*, 50:5–43, 2003.

M. Azizmalayeri, A. Abu-Hanna, and G. Cinà. Mitigating Overconfidence in Out-of-Distribution Detection by Capturing Extreme Activations. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024.

F. Bickford Smith, J. Kossen, E. Trollope, M. van der Wilk, A. Foster, and T. Rainforth. Rethinking Aleatoric and Epistemic Uncertainty. In *Workshop on Bayesian Decision-making and Uncertainty, 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

K. Chalupka, P. Perona, and F. Eberhardt. Visual Causal Feature Learning. In *UAI*, 2015.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, second edition, 2006.

S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

D. Draper. Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):45–70, 1995.

D. Dubois, H. Prade, and P. Smets. Representing partial ignorance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 26(3):361–377, 1996.

C. Eastwood, S. Singh, A. L. Nicolicioiu, and M. Vlastelica. Spuriosity Didn't Kill the Classifier: Using Invariant Predictions to Harness Spurious Features. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

W. Falcon. PyTorch Lightning, 2023.

M. Fellaji and F. Pennerath. The Epistemic Uncertainty Hole: An issue of Bayesian Neural Networks, 2024.

F. Friedrich, W. Stammer, P. Schramowski, and K. Kersting. A Typology for Exploring the Mitigation of Shortcut Behavior. *Nature Machine Intelligence*, 5:319–330, 2023.

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*, 56:1513–1589, 2023.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

C. Gruber, P. O. Schenk, M. Schierholz, F. Kreuter, and G. Kauermann. Sources of Uncertainty in Machine Learning – A Statisticians' View, 2023.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

P. Hofman, Y. Sale, and E. Hüllermeier. Quantifying Aleatoric and Epistemic Uncertainty: A Credal Approach. In *Structured Probabilistic Inference & Generative Modeling Workshop of ICML*, 2024.

N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel. Bayesian Active Learning for Classification and Preference Learning, 2011.

E. Hüllermeier and W. Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning*, 110:457–506, 2021.

P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. Wilson. What Are Bayesian Neural Network Posteriors Really Like? In *Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR 139*, 2021.

K. Jalaldoust, A. Bellot, and E. Bareinboim. Partial Transportability for Domain Generalization. Technical Report R-88, Columbia University, 2024.

A. Kendall and Y. Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. In *Proceedings of the IEEE*, 1998.

I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

T. P. Minka. Bayesian Model Averaging Is Not Model Combination, 2002.

B. Mlodozeniec, R. E. Turner, and D. Krueger. Implicitly Bayesian Prediction Rules in Deep Learning. In *Sixth Symposium on Advances in Approximate Bayesian Inference-Archival Track*, 2024.

B. Mucsányi, M. Kirchhof, and S. J. Oh. Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

V. Nagarajan, A. Andreassen, and B. Neyshabur. Understanding the Failure Modes of Out-of-Distribution Generalization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

M. Pagliardini, M. Jaggi, F. Fleuret, and S. P. Karimireddy. Agree to Disagree: Diversity through Disagreement for Better Transferability. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

J. Richens and T. Everitt. Robust agents learn causal world models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Y. Sale, P. Hofman, T. Löhr, L. Wimmer, T. Nagler, and E. Hüllermeier. Label-wise Aleatoric and Epistemic Uncertainty Quantification. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2024.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Towards Causal Representation Learning. *Proceedings of the IEEE*, 109(5), 2021.

K. Schweighofer, L. Aichberger, M. Ielanskyi, and S. Hochreiter. On Information-Theoretic Measures of Predictive Uncertainty, 2024.

L. Scimeca, A. Rubinstein, D. Teney, S. J. Oh, A. M. Nicolicioiu, and Y. Bengio. Mitigating Shortcut Learning with Diffusion Counterfactuals and Diverse Ensembles, 2024.

C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27: 379–423, 1948.

M. Shoja and E. S. Soofi. Uncertainty, information, and disagreement of economic forecasters. *Econometric Reviews*, 36(6-9):796–817, 2017.

F. B. Smith, A. Kirsch, S. Farquhar, Y. Gal, A. Foster, and T. Rainforth. Prediction-Oriented Bayesian Active Learning. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR 206, 2023.

D. Steinmann, F. Divo, M. Kraus, A. Wüst, L. Struppek, F. Friedrich, and K. Kersting. Navigating Shortcuts, Spurious Correlations, and Confounders: From Origins via Detection to Mitigation, 2024.

D. Teney, E. Abbasnejad, S. Lucey, and A. v. d. Hengel. Evading the Simplicity Bias: Training a Diverse Set of Models Discovers Solutions with Superior OOD Generalization. In *Proceedings of the 39th Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

A. M. Van Der Bles, S. Van Der Linden, A. L. J. Freeman, J. Mitchell, A. B. Galvao, L. Zaval, and D. J. Spiegelhalter. Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, 6(5):181870, 2019.

Y. Wald, A. Feder, D. Greenfeld, and U. Shalit. On Calibration and Out-of-domain Generalization. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

V. D. Wild, S. Ghalebikesabi, D. Sejdinovic, and J. Knoblauch. A Rigorous Link between Deep Ensembles and (Variational) Bayesian Methods. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

A. G. Wilson and P. Izmailov. Bayesian Deep Learning and a Probabilistic Perspective of Generalization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

L. Wimmer, Y. Sale, P. Hofman, B. Bischl, and E. Hüllermeier. Quantifying Aleatoric and Epistemic Uncertainty in Machine Learning: Are Conditional Entropy and Mutual Information Appropriate Measures? In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023.

A Appendix

A.1 Notation

In the following, we collect and define notational symbols used throughout the paper.

A.1.1 General

$\mathbb{R}$	set of real numbers
$\mathbb{N}$	set of natural numbers
$\mathcal{X} \subseteq \mathbb{R}^{d_x}, d_x \in \mathbb{N}$	feature (or input) space
$\mathcal{Y} \subseteq \mathbb{R}^{d_y}, d_y \in \mathbb{N}$	target (or label, output) space
$\mathcal{D} \in (\mathcal{X} \times \mathcal{Y})^n, n \in \mathbb{N}$	set of training data

A.1.2 Random Variables

$\mathbb{E}_q(\cdot)$	expectation w.r.t. distribution q
$D_{\mathrm{KL}}(q \\| \cdot)$	Kullback-Leibler divergence from distribution q
$H(\cdot)$	Shannon entropy
$I(\cdot;\cdot)$	mutual information

Shannon Entropy The *Shannon entropy* (Shannon, 1948) of a discrete random variable (RV) A with realizations in a sample space Ω is defined as:

$$H(A) = -\sum_{\omega \in \Omega} \omega \log \omega \;\in [0, \log |\Omega|], \tag{2}$$

where the logarithm is typically set to base 2 in accordance with an information-theoretic bit interpretation. Entropy captures the potential information gain from observing the realization of a RV. Consequently, it is minimal for a RV whose distribution is a Dirac measure, since the outcome is all but certain *a priori*, and maximal for uniformly distributed RVs (Cover and Thomas, 2006).

Mutual Information Entropic measures give rise to the *mutual information* between two RV:

$$\begin{aligned} I(A;B) &= H(A) - H(A|B) \\ &= D_{\mathrm{KL}}\left(p(A,B) \parallel p(A) \otimes p(B)\right) \\ &= \mathbb{E}_{p(B)}\left[D_{\mathrm{KL}}\left(p(A|B) \parallel p(A)\right)\right], \end{aligned} \tag{3}$$

with $\otimes$ denoting the outer product distribution. Mutual information is a measure of statistical independence and quantifies how much information can be gained about A by observing B, or *vice versa*. It vanishes at perfect independence, i.e., when the joint distribution factorizes into $p_A \otimes p_B$, and realizations from A do not decrease uncertainty over B. Alternatively, we can view $I(A;B)$ as the expected divergence between the conditional $p(A|B)$ and the marginal $p(A)$ that increases with more information in B about A (Cover and Thomas, 2006). This latter interpretation is particularly useful to understand the emphasis on disagreement between base learner and consensus prediction expressed in Eq. 1 and Eq. 10.

A.1.3 Uncertainty Decomposition

We largely adopt the notation of Hofman et al. (2024) in the following.

$\triangle_K, K \in \mathbb{N}$	$K-1$ simplex
$\mathbb{P}(\mathcal{Y})$	set of first-order probability distributions over $\mathcal{Y}$
$\mathcal{H} = \{h : \mathcal{X} \to \mathbb{P}(\mathcal{Y}) \mid h \text{ is of a certain functional form}\}$	space of probabilistic hypotheses
$Q : \mathcal{H} \to [0,1]$	second-order probability distribution
Y	RV of outcome labels
Θ	RV of first-order distributions
$M \in \mathbb{N}$	ensemble size

Bi-Level Uncertainty Representation For the classification case discussed here, we assume first-order label distributions $\boldsymbol{\theta} = p(Y|\cdot)$, , where Y denotes the random outcome variable, to be categorical with $K = |\mathcal{Y}| < \infty$ possible outcomes, equating the set $\mathbb{P}(\mathcal{Y})$ of such distributions with the $K-1$ simplex $\triangle_K = \{\boldsymbol{\theta} \in [0,1]^K : \|\boldsymbol{\theta}\|_1 = 1\}$.

Bayesian agents produce a probability distribution over the probabilistic prediction $\boldsymbol{\theta}$ (*posterior predictive density; PPD*) that is induced by the second-order distribution Q. Q assigns probabilities to hypotheses from $\mathcal{H}$. In the Bayesian paradigm, a prior belief $Q(\cdot)$ is updated to a posterior belief $Q(\cdot|\mathcal{D})$ after observing data, giving rise to the following PPD:

$$p(\boldsymbol{\theta}) = \int_{\mathcal{H}} \mathbf{1}_{h(\boldsymbol{x})=\boldsymbol{\theta}} \, \mathrm{d}Q(h|\mathcal{D}). \tag{4}$$

Here, $h(\boldsymbol{x}) \in \triangle_K$ models the ground-truth conditional density $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_K^*)^\top$ with $\theta_k^* = p(Y = k|\boldsymbol{x})$. We will sometimes omit the conditioning on $\boldsymbol{x}$ so as to not overload notation.

Bayesian Model Average The Bayesian paradigm further admits a first-order predictive distribution as an expectation over all possible models (hypotheses), yielding the *consensus prediction*

$$\bar{\boldsymbol{\theta}} = \int_{\mathcal{H}} h(\boldsymbol{x}) \, \mathrm{d}Q(h|\mathcal{D}). \tag{5}$$

In most practical problems, both Q and the integral in Eq. 5 are intractable. This issue is typically addressed by (unbiased) Monte Carlo integration over samples from Q (as obtained by some approximately Bayesian—e.g., sampling-based or variational—method; Andrieu et al., 2003). We specifically consider ensembles with M base learners (Wilson and Izmailov, 2020), leading to the following approximation:

$$\bar{\boldsymbol{\theta}} \approx \frac{1}{M} \sum_{m=1}^{M} h^{[m]}(\boldsymbol{x}). \tag{6}$$

Note that Eq. 6 is only a valid approximation of Eq. 5 if all ensemble members represent the same structural form of hypothesis (Minka, 2002). This is the case for deep ensembles (Lakshminarayanan et al., 2017), where base learners differ solely by parameterization as a consequence of random weight initialization and stochastic training elements.

Entropy Decomposition With Θ denoting the RV whose realizations are distributions $\boldsymbol{\theta} \in \triangle_K$, we can derive the components of predictive uncertainty as

$$\underbrace{H(Y)}_{\text{TU}} = \underbrace{H(Y|\Theta)}_{\text{AU}} + \underbrace{I(Y;\Theta)}_{\text{EU}}. \tag{7}$$

(TU) The *total* uncertainty of a prediction obtained via Bayesian model averaging (Eq. 5) is quantified via Shannon entropy (Eq. 2) and defined as

$$H(Y) = H\left(\mathbb{E}_Q\left[Y|\Theta\right]\right) = H(\bar{\boldsymbol{\theta}}) = -\sum_{k=1}^{K} \bar{\theta}_k \log \bar{\theta}_k. \tag{8}$$

The more $\bar{\boldsymbol{\theta}}$ concentrates on a single outcome (pushing it toward one of the simplex corners), the lower its corresponding uncertainty.

(AU) Similarly, we obtain *aleatoric* uncertainty as the *conditional entropy* of the outcome:

$$H(Y|\Theta) = \mathbb{E}_Q \left[H(Y|\Theta) \right] = - \int p(\boldsymbol{\theta}) H(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}. \tag{9}$$

(EU) The *epistemic* component emerges as the residual quantity from the additive decomposition of Eq. 8, which amounts to the mutual information between Y and Θ:

$$\begin{aligned} I(Y;\Theta) &= H(Y) - H(Y|\Theta) \\ &= \mathbb{E}_Q \left[D_{\mathrm{KL}}(p(Y|\boldsymbol{\theta}) \parallel p(Y) \right] \\ &= \mathbb{E}_Q \left[D_{\mathrm{KL}} \left(p(Y|\boldsymbol{\theta}) \parallel \bar{\boldsymbol{\theta}} \right) \right]. \end{aligned} \tag{10}$$

A.2 Experiments

A.2.1 Experimental Details

Datasets We consider tasks based on MNIST (LeCun et al., 1998). Training sets comprise 10k images with balanced classes:

- **MNIST3.** 3-class MNIST version where original classes are pooled into classes $\{0, 1, 2\}$, consisting of digits 1-3, 4-6, 7-9, respectively (leaving class 0 as OOD test data).
- **CMNIST3.** Version of MNIST3 with global shortcut coloring digits 1-3 in red, digits 4-6 in green, and digits 7-9 in blue (similar experiments are conducted in, e.g., Jalaldoust et al., 2024).
- **PMNIST3.** Version of MNIST3 with local shortcut adding a colored 1×1 patch in the top left of each image; digits 1-3: red, digits 4-6: green, digits 7-9: blue (similar experiments are conducted in, e.g., Adnan et al., 2022).
- **MNIST0.** Class 0 from original MNIST data, used as OOD data.

We vary the shortcut strength by modifying $s\%$ of images (e.g., CMNIST3 with $s = 50$ is created by coloring 50% of MNIST3 images and leaving the rest black-and-white).

Models We use deep ensembles (Lakshminarayanan et al., 2017) of small NNs with one convolutional layer (16 filters of size 3×3), followed by max-pooling and two fully-connected layers (dimensions $16 \cdot 196$ and 128, respectively), ReLU activations in the hidden layers, and softmax activation in the final layer.

Training We train our models for 25 epochs with AdamW optimization (Loshchilov and Hutter, 2019), an initial learning rate of 0.001 that is reduced by a factor of 0.1 if it plateaus for 10 consecutive epochs, and weight decay of 0.01. Batch size is set to 128.

Software Our code is mainly based on `PyTorch` (Paszke et al., 2019) and `PyTorch Lightning` (Falcon, 2023). We performed all experiments on CPU and will make our code available upon publication.

A.2.2 Further Results

For all results, we normalize the uncertainty components to values in $[0, 1]$ (using the upper bound on Shannon entropy, Eq. 2, given by the number of possible outcomes).

More Epochs Fig. 3 shows results on MNIST0 test data for some more training epochs in addition to Fig. 2. We observe that model weights and their magnitude changing over the course of training also affect the depicted uncertainty estimates.

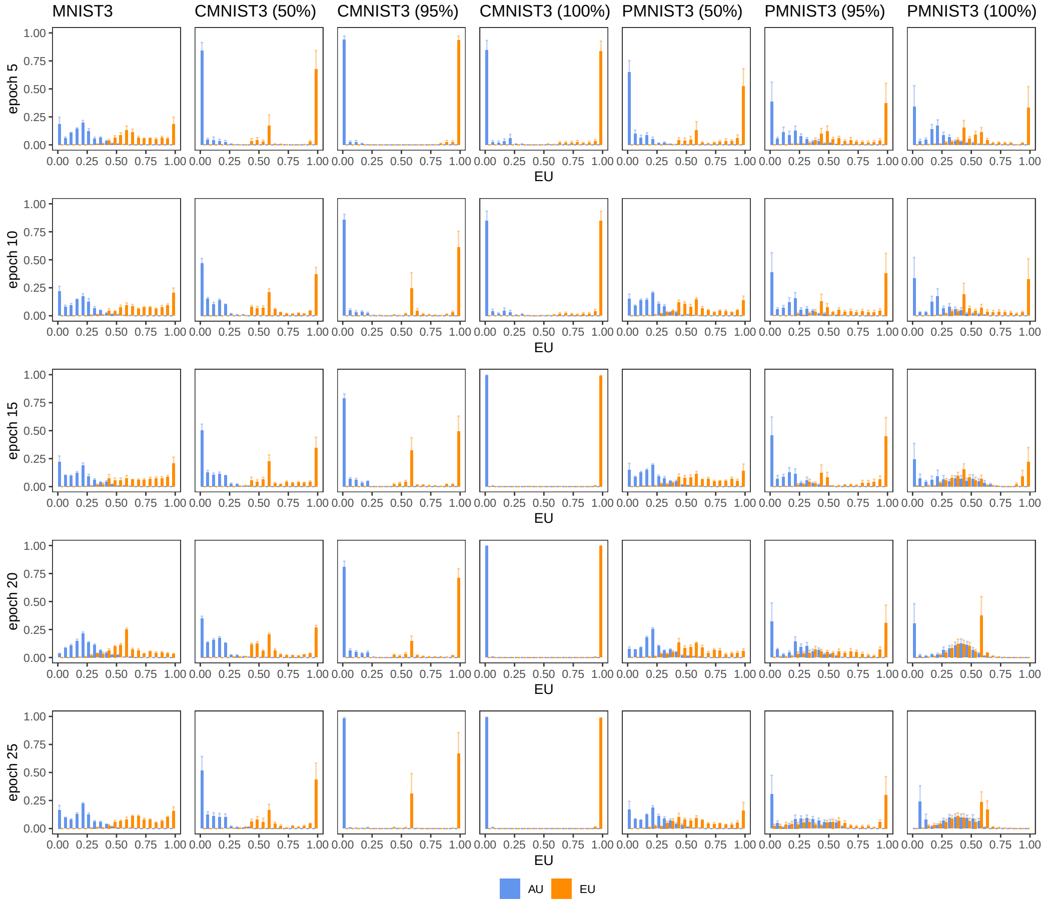

Figure 3: *More epochs.* Uncertainty estimates (20% least-confident predictions) based on data with varying s; ensemble size 3; average over 3 independent runs (error bars: ± 1 standard error).

Effect of Shortcut Strength For CMNIST3, some effect is visible even when only half of the training images are colored (column 2 in Fig. 3). PMNIST3 (column 5) produces uncertainty estimates as diffuse as for the MNIST3 (column 1) data without any shortcut; disagreement-seeking behavior only occurs for strong shortcuts. In the case of a perfect spurious correlation, the CMNIST3-trained learner (column 4) settles for full confidence and full disagreement on virtually all test samples. PMNIST3 (column 7) exhibits a sort of model collapse, where hypotheses start to agree more in later epochs. Stronger correlations also lead to lower accuracy[9] (Tab. 1) as the models come to rely on the shortcut and pick up little of the stable pattern. For instance, the learner trained on CMNIST3 achieves 95% accuracy (epoch 25) when exposed to 50% shortcut strength, nearly on par with MNIST3, but makes heavy use of the 100% coloring cue and deteriorates to 41% accuracy in the strong-shortcut scenario.

Larger Ensemble Size When we increase the ensemble size to 5 (Fig. 4, as opposed to size 3 for the results reported in Fig. 2, with otherwise identical settings), we observe a similar tendency for CMNIST3 toward strong disagreement. The PMNIST3 training still provokes more conflict than shortcut-free MNIST3 but to a lesser degree than CMNIST3. In general, with growing ensemble size, the probability of some members converging to the same predictions rises. Note that it is no longer possible for all hypotheses to settle on completely conflicting predictions: Since the number

[9] Accuracy is calculated with MNIST3 as a test set (i.e., images from the same classes as in the training data, but free of shortcuts—accuracy on the never-seen OOD class would be consistently 0).

Epoch	MNIST3	CMNIST3 50	CMNIST3 95	CMNIST3 100	PMNIST3 50	PMNIST3 95	PMNIST3 100
5	0.84 (0.05)	0.86 (0.01)	0.62 (0.02)	0.44 (0.04)	0.72 (0.11)	0.68 (0.11)	0.57 (0.06)
10	0.95 (0.00)	0.90 (0.00)	0.61 (0.01)	0.44 (0.04)	0.86 (0.06)	0.67 (0.11)	0.57 (0.06)
15	0.96 (0.00)	0.93 (0.00)	0.72 (0.01)	0.41 (0.02)	0.87 (0.06)	0.67 (0.11)	0.56 (0.06)
20	0.97 (0.00)	0.95 (0.00)	0.67 (0.01)	0.41 (0.02)	0.86 (0.06)	0.67 (0.11)	0.68 (0.11)
25	0.97 (0.00)	0.95 (0.00)	0.67 (0.03)	0.41 (0.03)	0.87 (0.06)	0.67 (0.11)	0.68 (0.11)

Table 1: Accuracy and corresponding standard errors over 3 independent runs; ensemble size 3.

of ensemble members is now larger than (and not a multiple of) the number of classes, the maximum EU value of 1 is not attainable in this setting[10].

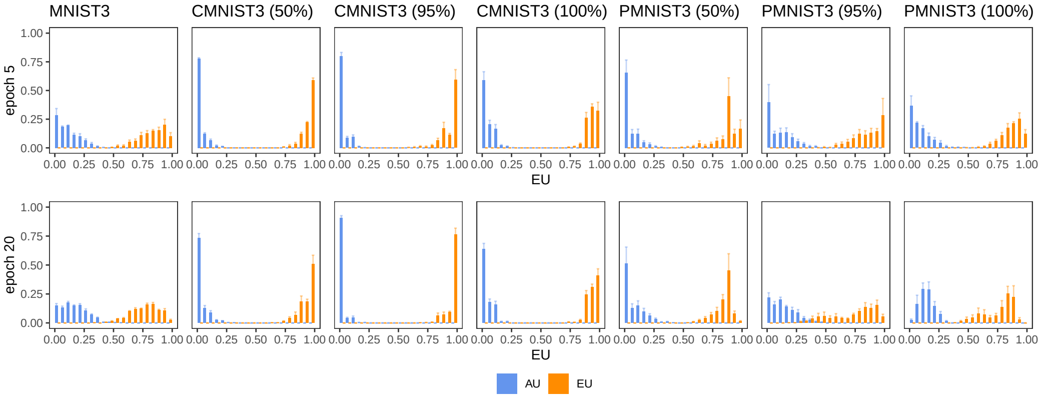

Figure 4: *Larger ensemble.* Uncertainty estimates (20% least-confident predictions) based on data with varying s; ensemble size 5; average over 3 independent runs (error bars: ± 1 standard error).

[10] EU values may still end up in the highest bin of (0.95, 1]: 5 predictions for some class at full confidence each, with at most two members agreeing at a time will produce some perturbation of a (0.2, 0.4, 0.4) class probability distribution. This distribution has 0.96 entropy (TU), while AU, as the average entropy over one-hot probability vectors, is 0. Per the additivity constraint of the entropy decomposition, this leaves EU = TU - AU at 0.96, and thus in the last bin.