Uncertainty Definitions and Bayesian Representation (AU, EU, TU)

Expert Response

November 23, 2025

Abstract

This document provides the definitions of Aleatoric Uncertainty (AU), Epistemic Uncertainty (EU), and Total Uncertainty (TU), detailing their theoretical Bayesian representation and practical modeling, particularly within the context of Explainable Artificial Intelligence (XAI).

1 Definitions of AU, EU, and TU

Uncertainty quantification (UQ) aims to provide a reliable representation of a model's limitations by distinguishing between two fundamental sources of uncertainty: Aleatoric and Epistemic Uncertainty [8, 9].

Table 1: Definitions of Uncertainty Types

Term	Full Name	Definition and Characteristics
TU	**Total Uncertainty**	This is the aggregate predictive uncertainty, encompassing both sources of error [8, 9].
AU	**Aleatoric Uncertainty** (Data Uncertainty)	Arises from the inherent stochasticity, noise, or randomness within the data generating process itself [8, 9]. It reflects the non-deterministic relationship between input and output [8, 9]. It is generally considered **irreducible** by observing more examples of the phenomenon, although gathering more features might alleviate it [8, 9].

Table 1: Definitions of Uncertainty Types

Term	Full Name	Definition and Characteristics
EU	**Epistemic Uncertainty** (Model Uncertainty)	Stems from the learning algorithm's ignorance of the true underlying model [8, 9]. It relates to **model multiplicity**, where multiple models can achieve similar performance [8]. It is potentially **reducible** by acquiring additional data or observations, thereby training the model more reliably [8, 9].

2 Theoretical Representation using Bayesian Methods

Bayesian learning provides the fundamental framework for representing and decomposing uncertainty by assuming a probability distribution over model parameters Θ, leading to predictive distributions $p(y|x)$ [8, 9].

The decomposition of Total Uncertainty (TU) relies on quantifying the uncertainty based on the shape of the output distribution (first-order distribution) and the distribution over those outputs (second-order distribution) [8, 9].

2.1 Classification (Using Shannon Entropy, H_S)

When measuring uncertainty for classification using Shannon Entropy, $H_S[p(y|x)]$, the theoretical decomposition into Aleatoric Uncertainty (AUS) and Epistemic Uncertainty (EUS) is based on the relationship between model variance and intrinsic data noise. The total uncertainty is decomposed as follows[8, 9]:

$$\mathrm{TU}(x) = H_S[p(y|x)] = \underbrace{\mathbb{E}_\Theta[H_S[p(y|x,\theta)]]}_{\text{Aleatoric Uncertainty (AUS)}} \underbrace{+\,\mathbb{E}_{p(y,\theta|x)}\left[\log\left(\frac{p(y \mid x)\,p(\theta \mid x)}{p(y,\theta \mid x)}\right)\right]}_{\text{Epistemic Uncertainty (EUS)}}$$

- **AUS** represents the expected entropy of the predictive distribution over the posterior distribution of the parameters. Since a fixed set of parameters θ eliminates model uncertainty, the remaining uncertainty is purely due to irreducible data noise [8, 9].

- **EUS** captures the uncertainty resulting from ignorance about the true parameters Θ [8, 9]. It is calculated as the difference: $EUS(x) = H_S(x) - AUS(x)$ [8, 9].

2.2 Regression (Using Variance, H_V)

For regression tasks, total uncertainty is quantified using variance, $H_V(x) = V_Y[y|x]$. The decomposition is based on the law of total variance:

$$H_V(x) = \underbrace{E_\Theta[V_Y[y|x,\theta]]}_{\text{Aleatoric Uncertainty (AUV)}} + \underbrace{V_\Theta[E_Y[y|x,\theta]]}_{\text{Epistemic Uncertainty (EUV)}}$$

- **AUV** (Aleatoric Variance) measures the expected variance of the outcome given the model parameters, reflecting irreducible noise [8, 9].
- **EUV** (Epistemic Variance) measures the variance of the mean outputs across different sampled parameters, reflecting the model's uncertainty about its optimal parameters [8, 9].

3 Practical Modelling using Bayesian Neural Networks

In practice, full Bayesian inference is often intractable, so approximations, typically using **Bayesian Neural Networks (BNNs)** or ensembles thereof, are employed to quantify uncertainty [8, 9, 7]. This involves drawing a finite sample of weights $\{\theta_1, \ldots, \theta_n\}$ from the approximate posterior distribution [8].

3.1 Practical Quantification using Finite Samples

When dealing with a finite sample of model predictions $p(y|\theta_i, x)$, the uncertainties are computed as follows:

Table 2: Practical Uncertainty Formulas (Finite Sample)

Uncertainty Type	**Formula (Classification using Entropy / Regression using Variance)**
Aleatoric Uncertainty (AU)	$\text{AUS}(x) = -\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{y\in Y} p(y|\theta_i,x)\log_2 p(y|\theta_i,x)\right)$ (Average Entropy) $\text{AUV}(x) = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$ (Average Variance of output distribution) [8, 9]
Epistemic Uncertainty (EU)	$\text{EUS}(x) = \text{TUS}(x) - \text{AUS}(x)$ $\text{EUV}(x) = \frac{1}{n}\sum_{i=1}^{n}(\mu_i - \mu^*)^2$ (Variance of the predicted means) [8, 9]
Total Uncertainty (TU)	TUS is the entropy computed from the averaged probability distribution $\frac{1}{n}\sum_{i=1}^{n} p(y|\theta_i, x)$. $\text{TUV}(x) = \text{AUV}(x) + \text{EUV}(x)$.

3.2 Application in Counterfactual Explanations (CE)

Uncertainty quantification provides a unifying framework for generating Counterfactual Explanations (CE) by linking uncertainty measures to desirable CE properties [8, 9]. Approaches like CLUE (Counterfactual Latent Uncertainty Explanations) embed uncertainty minimization into the explanation generation objective [10, 9, 7].

- **Minimizing AU for Validity and Discriminativeness:** Low AU means the inherent data noise is low, ensuring the resulting counterfactual x^{CF} is confidently classified into the target class τ (Validity) and is unambiguously separable from other classes (Discriminativeness) [8, 9, 9].
- **Minimizing EU for Plausibility and Feasibility:** Low EU means the model is certain about its prediction at x^{CF}. Points far from the training data manifold exhibit high EU, so minimizing EU guides the counterfactual toward the observed data distribution (Plausibility/Feasibility) [8, 9].

The objective functions for finding CEs often utilize optimization (OPT) strategies [4, 6] or Multi-Objective Optimization (MOO) [2, 5, 6]. A general optimization loss function $\mathcal{L}(x^{CF})$ often combines prediction certainty (Validity) and distance [1]:

$$\mathcal{L}(x^{CF}) = \underbrace{\mathcal{H}(y|\mu_\Theta(x|z))}_{\text{Uncertainty/Validity measure}} + \underbrace{d(\mu_\Theta(x|z), x_0)}_{\text{Distance metric}}$$

In multi-objective optimization, a counterfactual search problem aims to minimize multiple functions simultaneously [5, 2]:

$$\min_{\theta \in C} F(\theta), \quad \text{where} \quad F(\theta) = (F_1(\theta), F_2(\theta), \dots, F_L(\theta))$$

where $F_L(\theta)$ represents L objectives, such as minimizing cost (distance) and ensuring high predictive validation, potentially across multiple models (model multiplicity) [5].

References

[1] Andrei Buliga, Chiara Di Francescomarino, Chiara Ghidini, Marco Montali, and Massimiliano Ronzani. Generating Counterfactual Explanations Under Temporal Constraints. *arXiv preprint arXiv:2502.10418*, 2025. See Eq. (7) for the general loss function structure including compliance.

[2] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-Objective Counterfactual Explanations. In *Parallel Problem Solving from Nature – PPSN XVI, 16th International Conference, PPSN 2020, Leiden, The Netherlands, September 5-9, 2020, Proceedings, Part I*, volume 12269, pages 448–469. Springer, 2020. Also available as arXiv:2004.11165 [stat.ML] [3].

[3] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-Objective Counterfactual Explanations. *CoRR*, abs/2004.11165, 2020.

[4] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38:2770–2824, 2024.

[5] Keita Kinjo. Robust Counterfactual Explanations under Model Multiplicity Using Multi-Objective Optimization. *arXiv preprint arXiv:2501.05795*, 2025. See Section 2.2 and optimization objective.

[6] Marcos M Raimundo, Luis Gustavo Nonato, and Jorge Poco. Mining pareto-optimal counterfactual antecedents with a branch-and-bound model-agnostic algorithm. In *Preprint*, 2025. Relevant for multi-objective optimization concept (MOC).

[7] Lisa Schut, Oscar Key, Rory McGrath, Luca Costabello, Bogdan Sacaleanu, Medb Corcoran, and Yarin Gal. Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleatoric uncertainties. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130, pages 1756–1764. PMLR, 2021.

[8] Kacper Sokol and Eyke Hüllermeier. All You Need for Counterfactual Explainability Is Principled and Reliable Estimate of Aleatoric and Epistemic Uncertainty. *CoRR*, abs/2502.17007, 2025. Foundational paper on uncertainty decomposition for CE.

[9] Santo Maria Amado Rocco Thies. Uncertainty as a Unifying Framework for Counterfactual Explanations, 2025. See Chapters 3, 4, and Appendix A.

[10] Arnaud Van Looveren and Janis Klaise. Interpretable Counterfactual Explanations Guided by Prototypes. In *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II*, volume 12976, pages 650–665. Springer, 2021.