Uncertainty in AI Conference

Not Just How Much, But Where: Decomposing Epistemic Uncertainty into Per-Class Contributions

Mame Diarra Toure, David A. Stephens
Department of Mathematics and Statistics, McGill University

McGill

THE PROBLEM

A single score hides the kind of mistake the model is about to make.

- Mutual information says how uncertain the model is, but not which classes are driving that uncertainty.
- In asymmetric tasks, that missing structure matters: benign-vs-benign confusion is not the same as benign-vs-critical confusion.

KEY IDEA: Decompose MI into class-specific epistemic contributions.

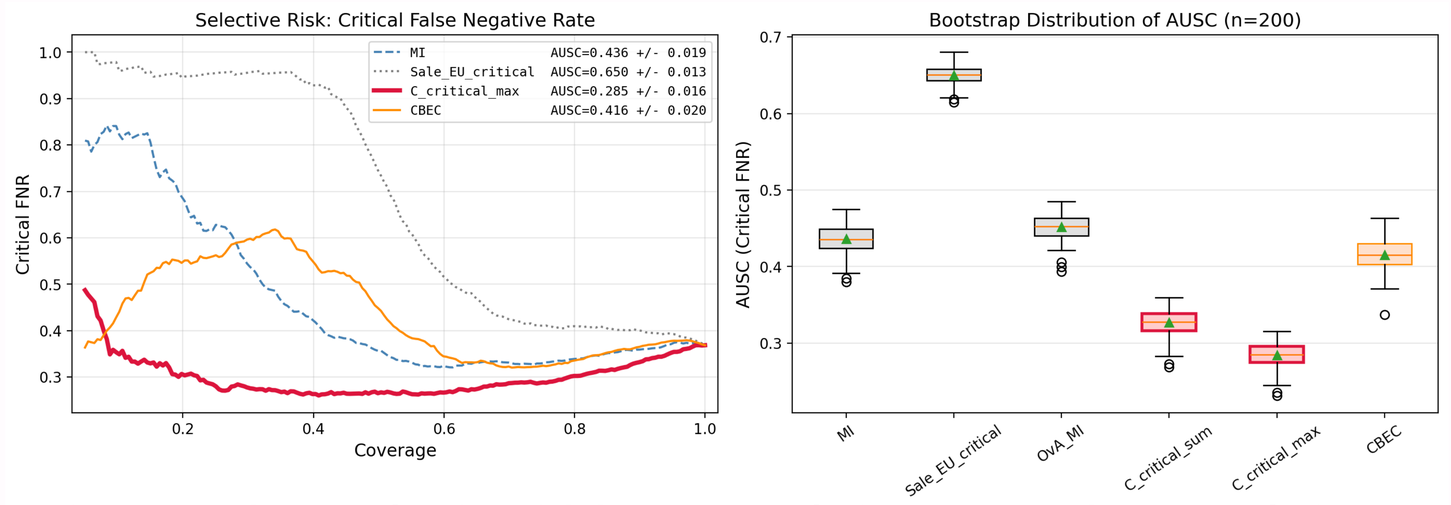
$$C_k(x) = \mathrm{Var}[p_k](x) / (2\,\mu_k(x))$$

Each class gets its own epistemic contribution, and the sum stays tied to scalar MI. The 1 / mu_k factor corrects boundary suppression and and makes C_k comparable across rare and common classes..

STORY OF THE PAPER

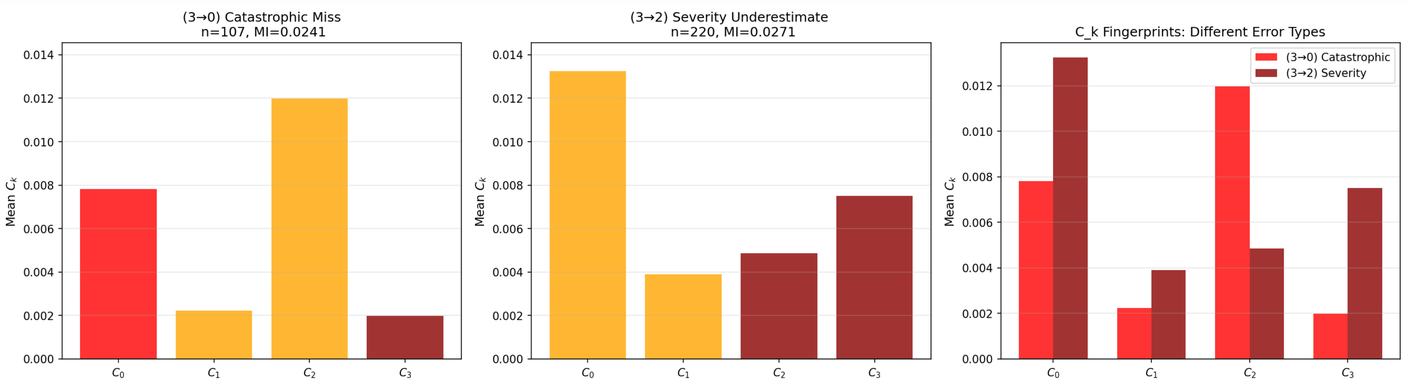
1. **Localize** — The decomposition turns one scalar into a class-by-class risk map.
2. **Explain** — Show that similar MI can correspond to very different failure modes
3. **Warn** — Show that good metrics still fail when the posterior is unhealthy, especially under transfer learning

Critical-class uncertainty yields safer deferral.



On diabetic retinopathy, C_crit_max reduces selective risk by 34.7% relative to MI and 56.2% relative to variance baselines. It defers where critical confusion lives, not just where uncertainty is high.

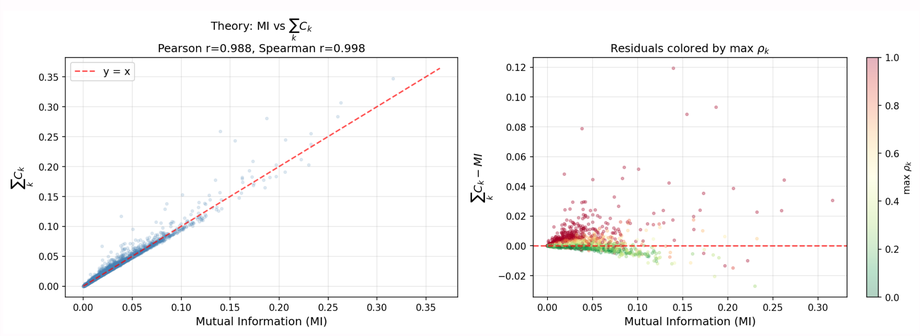
Same MI, different errors



Catastrophic misses and severity underestimates have nearly identical MI (0.024 vs 0.027 nats) but very different C_k fingerprints. The decomposition exposes which class drives the error and suggests different interventions.

TRUST CHECK

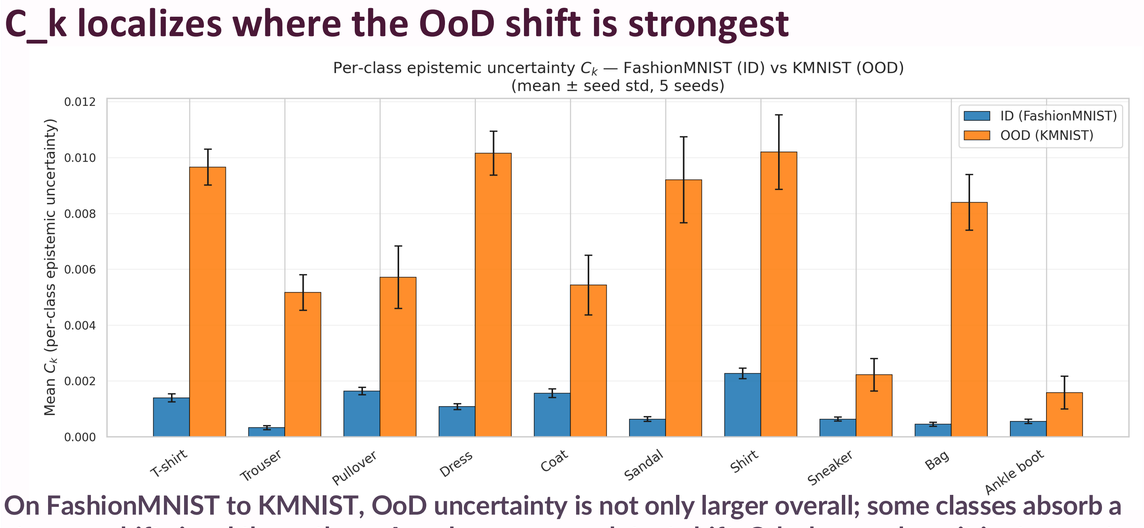
The class-wise decomposition stays anchored to MI.



Across 7,948 DR test samples, sum_k C_k tracks exact MI closely (Pearson 0.988, Spearman 0.998). The method adds localization without drifting far from the standard scalar uncertainty measure.

BEYOND DR

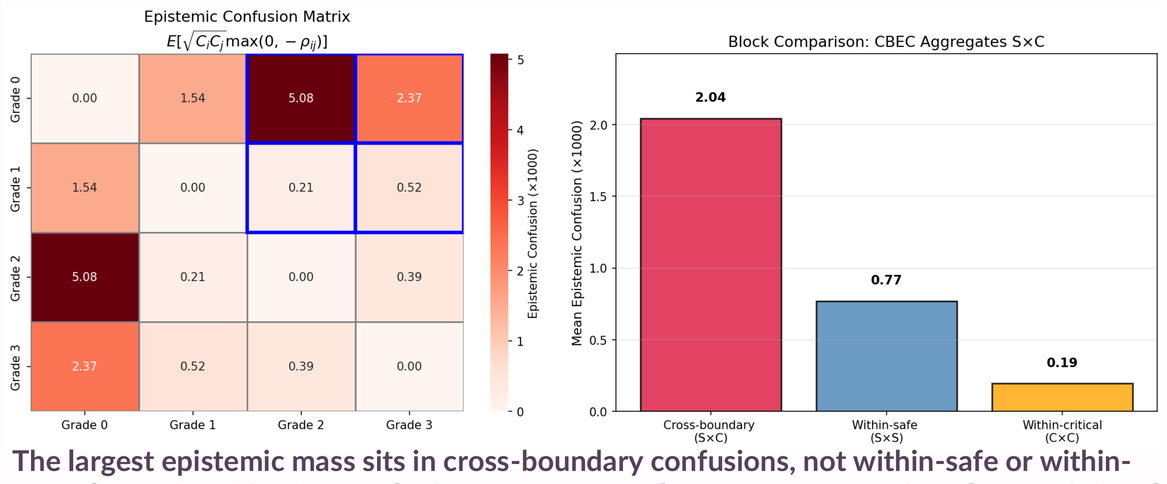
C_k localizes where the OoD shift is strongest



On FashionMNIST to KMNIST, OoD uncertainty is not only larger overall; some classes absorb a stronger shift signal than others. A scalar score can detect shift. C_k shows where it is strongest.

RISK STRUCTURE

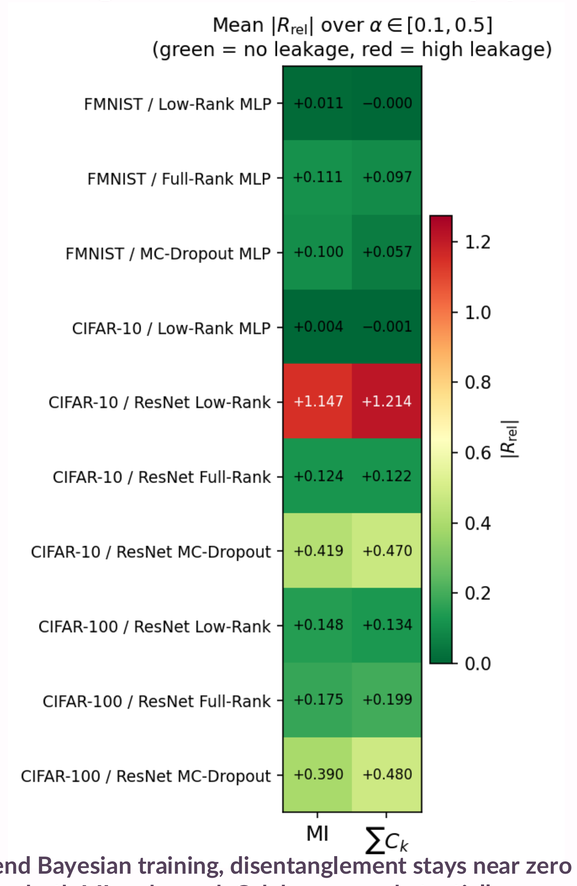
Uncertainty concentrates at the safe-critical boundary.



The largest epistemic mass sits in cross-boundary confusions, not within-safe or within-critical mistakes. That is exactly the structure a scalar score erases and a selective deferral policy needs.

FINAL LESSON

Transfer learning can break uncertainty quality.



Under end-to-end Bayesian training, disentanglement stays near zero leakage. Under transfer learning, both MI and sum_k C_k become substantially more entangled. A better score helps, but it cannot fully rescue an unhealthy posterior.

ENDING MESSAGE

Better metrics still need healthy uncertainty estimates.

0.004 CIFAR-10 low-rank from scratch, MI leakage	**-0.001** CIFAR-10 low-rank from scratch, sum_k C_k leakage	**1.147** CIFAR-10 ResNet transfer, MI leakage	**1.214** CIFAR-10 ResNet transfer, sum_k C_k leakage

- Use C_crit_max when specific classes are safety-critical and the goal is targeted deferral.
- Use the per-class OoD view when you want to know where the shift is strongest, not only whether it exists.
- Use sum_k C_k when the shift could land anywhere and you need a single monitoring score.
- Do not assume transfer learning gives trustworthy epistemic uncertainty; disentanglement can deteriorate even when the Bayesian head is unchanged.
- How uncertainty is propagated through the network matters as much as how it is measured.

MI tells you how much uncertainty there is. C_k tells you where it sits, and our disentanglement results show that this picture is clearest when epistemic and aleatoric uncertainty are well separated.

Bottom line

Not just how uncertain, but where the uncertainty lives.

Per-class epistemic structure changes the decision. It improves critical-class selective prediction, distinguishes failure modes with the same MI, localizes where OoD shift is strongest, and exposes when transfer learning yields unhealthy uncertainty estimates.