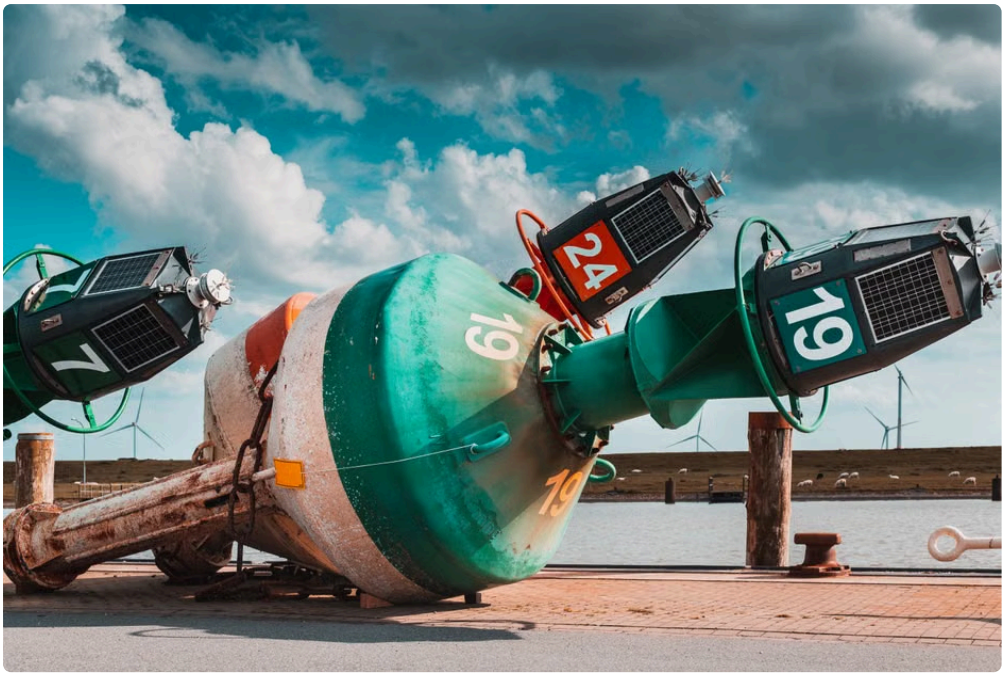
```html

Risk assessment models are increasingly used in criminal justice to inform decisions such as pretrial detention, sentencing, and parole. These models promise to bring objectivity and data-driven rigor to a historically biased and opaque decision-making process. However, they also introduce new concerns about fairness, accuracy, and unintended harms—especially when deployed across **protected groups** like racial minorities or economically disadvantaged populations.

One underappreciated lens for examining risk assessment models is through their *disagreement*—instances where predictions diverge across model versions, algorithmic ensembles, or between models and human experts. Disagreement is more than just noise; it can be a **high-signal indicator** of risk, edge cases, and potential failure points. In this article, we explore why auditing **disagreement** and related metrics such as *disagreement rate* and *predictive entropy* is critical for identifying data gaps, distribution shifts, and fairness risks in criminal justice risk models.

## Why Disagreement Matters in Risk Assessment Models

At first glance, disagreement might seem like a mere artifact of model instability or uncertainty. Yet, disagreement often marks cases that are challenging for the model—cases where predictions are less confident or more variable. These instances may:

- Represent **edge cases** outside the training distribution
- Highlight **data gaps** or insufficient subgroup representation
- Reveal **objective mismatches** between stakeholders' priorities and model loss functions
- Indicate higher **potential for disparate impact** on protected groups

Understanding disagreement is essential for auditors seeking to ensure that risk assessment models do not perpetuate or exacerbate bias and that they operate robustly across diverse populations.

## Key Tools for Measuring Disagreement: Disagreement Rate and Predictive Entropy

Two insightful and interpretable metrics capture disagreement:

### 1. Disagreement Rate

Here's what kills me: the disagreement rate quantifies how often predictions differ across multiple models or model iterations for the same inputs. For example, if three versions of a risk model provide binary scores (high-risk vs. low-risk), the disagreement rate is the proportion of cases where these models do not unanimously agree.

**Why it matters:** High disagreement rates often correlate with uncertainty and risk in decision outcomes. Monitoring these rates helps identify:

- Populations or subgroups where the model is less confident or consistent
- Changes over time indicating *distribution shifts*
- Areas with limited data or ambiguous signals

### 2. Predictive Entropy
```

Predictive entropy measures uncertainty in the model's predicted probability distribution—essentially a gauge of "confidence." For a risk assessment model outputting probability scores (e.g., probability of recidivism), higher entropy signals more uncertainty or spread in prediction, while low entropy indicates confident predictions near 0 or 1.

Formally, predictive entropy H is calculated as:

H(p) = - ∑ p_i log(p_i) https://reportz.io/ai/when-models-disagree-what-contradictions-reveal-that-a-single-ai-would-miss/

where p_i are predicted probabilities for each outcome.



Why it matters: Predictive entropy helps detect ambiguous cases that could require human review or flag potential model blind spots.

Disagreement as a High-Signal Risk Indicator

From my experience building ML systems for sensitive domains, the question is always: *what happens on the worst day in production?* Disagreement identifies those "worst day" inputs—cases where models are prone to error and disagreement with stakeholders.

When disagreement spikes, it is a red flag for:

1. **Uncertainty in individual predictions**—the model struggles to assign a clear risk level.
2. **Potential for harmful decision outcomes**—mistakes are more likely, possibly leading to wrongful detention or unwarranted leniency.
3. **Invariant behavior masking subgroup disparities**—disagreement may disproportionately emerge in protected groups, suggesting fairness concerns.

For example, instances where different models disagree on whether a defendant poses high risk might correspond to individuals with complex or uncommon profiles, which existing data fails to capture well. Tracking disagreement should be part of any robust risk model monitoring framework.

Edge Cases and Distribution Shift: Where Data Gaps Appear

Disagreement is often highest in regions of the feature space where:

- Training data is sparse
- Feature distributions shift over time (e.g., new crime patterns)
- Demographic composition changes in protected groups

These *edge cases* represent outliers or novel examples that the model has never encountered. Without auditing disagreement, these gaps remain invisible until failure occurs in production—sometimes with damaging consequences.

Consider a model trained primarily on urban populations that then encounters rural defendants with different socioeconomic factors. If disagreement and predictive entropy spike for these cases, it signals a need to retrain or augment data to regain reliable performance.

Data Gaps and Subgroup Coverage: Mitigating Disparate Impact through Auditing

Disparate impact occurs when risk models systematically disadvantage particular protected groups—often minorities—in a way not justified by actual risk. Disagreement auditing can illuminate hidden biases by revealing that model confidence and agreement vary significantly across subgroups.

To audit for this, one should:

1. Calculate disagreement rate and predictive entropy **within each protected group**
2. Compare performance metrics to detect uneven coverage or accuracy gaps
3. Investigate cases with high disagreement and their demographic attributes

Failure to do so risks perpetuating "silent" bias masked by aggregate test set accuracy metrics. Instead, subgroup-level disagreement audits drive targeted data collection or model adjustments to improve equity.

Objective Mismatch and Loss Function Tradeoffs

Another subtle cause of disagreement lies in the mismatch between the *optimization objective* that training algorithms pursue and the multiple, sometimes conflicting, objectives of stakeholders.

For instance, a commonly used loss function like cross-entropy minimizes overall error but does not explicitly account for minimizing disparate impact or promoting fairness. Different model configurations optimized for different criteria might disagree strongly on individual cases.

Auditing disagreement can expose such tradeoffs and force teams to:

- Explicitly define risk thresholds tied to real-world costs of false positives vs. false negatives
- Incorporate domain-specific fairness constraints in model training
- Balance between predictive accuracy and equitable treatment across groups

In other words, disagreement is often the symptom of deeper tensions in loss function design and stakeholder priorities.

Best Practices for Auditing Disagreement in Criminal Justice Risk Models

Based on lessons learned from launching and monitoring decision systems under risk, here are recommended steps for effective disagreement audits:

1. **Collect multiple model versions and/or ensemble outputs** to compute disagreement rates.
2. **Compute predictive entropy** for all predictions to assess confidence distribution.
3. **Stratify disagreement and entropy by protected group and relevant subpopulations.**
4. **Monitor changes over time** to detect distribution shifts and emerging risk zones.
5. **Investigate edge cases** with high disagreement through manual review or domain expert input.
6. **Align threshold settings to cost-based decision frameworks** rather than arbitrary cutoffs.
7. **Report disagreement metrics alongside accuracy and fairness metrics** to provide a fuller picture.

Things Accuracy Hides: Why Disagreement is Indispensable

Accuracy numbers alone lie. They smooth over population heterogeneity, ignore distribution shifts, and hide uncertainty. Here's a quick checklist of "things accuracy hides" that disagreement catches:

- Unequal confidence on protected groups
- Ambiguous cases with high error risk
- Emergent data regime changes

- Loss function blind spots causing misalignment with domain costs
- Potential overfitting to majority populations

Auditors and practitioners who focus only on aggregate test accuracy are vulnerable to “silent failures” that undermine justice and trust.

Conclusion: Disagreement Audits Are Not Optional

In criminal justice risk assessment, where decisions substantially impact human lives and freedoms, auditing disagreement is a practical and ethical imperative. Disagreement rate and predictive entropy provide easily computed, actionable signals that surface hidden risk, fairness gaps, and robustness issues.

More than just a debugging tool, disagreement auditing aligns risk model development with transparent, equitable, and cost-aware decision-making. As ML practitioners and stakeholders, our focus should be on answering:

What happens on the worst day in production? And how can disagreement help us predict and prevent it?

Only then can we hope to build risk assessment models that truly serve justice rather than merely creating new inequities beneath a veneer of objectivity.

```
```