**Ethical Reflection**

**Addressing Biases with IBM AI Fairness 360 (AIF360)**

The IBM AI Fairness 360 toolkit provides comprehensive algorithms to detect and mitigate bias throughout the ML pipeline:

1. **Bias Detection**:

**Metric**: Use Classification Metric to calculate:

* Disparate impact ratio across demographic groups
* Equal opportunity difference (e.g., recall parity)
* Average odds difference

1. **Pre-processing Mitigation**:

**Technique**: Use Reweighing to adjust sample weights

* Balances influence of underrepresented groups
* Example: Increase weight of darker skin tone images

1. **In-processing Mitigation**:

**Technique**: Apply adversarial debiasing during training

* Uses adversarial networks to remove protected attributes from predictions

1. **Post-processing Mitigation**:

**Technique**: Implement EqualizedOddsPostprocessing

* Adjusts decision thresholds per subgroup
* Ensures equal FPR/FNR across demographics

**Ethical Implementation Framework**

1. **Continuous Monitoring**:

* Track model performance across subgroups monthly

1. **Human-in-the-Loop Design**:
   * Maintain clinician oversight for high-risk predictions
   * Implement override mechanisms when:

* Confidence scores are borderline (e.g., 45-55%)
* Cases involve rare demographics

1. **Transparency Measures**:

Generate explainability reports using:

* SHAP values: shap.DeepExplainer(model, X\_train)
* LIME: lime\_image.LimeImageExplainer()

Example output: Visual heatmaps showing features driving predictions

1. **Compensation Mechanisms**:
   * For underserved populations:

* Prioritize collection of missing data
* Allocate 30% of R&D budget to targeted data acquisition
* Partner with community health centers