# Information Theoretic Fairness

Agus Sudjianto, Ph.D
Center for Trustworthy AI through Model Risk Management
University of Noth Carolina Charlotte

# Fragmented Fairness Metrics

- Fairness metrics were defined ad hoc, based on moral intuitions or statistical heuristics. They lack a common mathematical base.
- Fragmented and inconsistent, leading to conflicting criteria that hinder practical adoption in real-world applications.
- We need a unified framework.

| Fairness Metric | Independence Condition | Intuition | What It Protects | Typical Violation Example |
|---|---|---|---|---|
| **Demographic Parity (DP)** | $\hat{Y} \perp\!\!\!\perp A$ | Predictions should be independent of group membership. | Outcome rate parity | Different approval rates |
| **Equalized Opportunity (EO)** | $\hat{Y} \perp\!\!\!\perp A \mid (Y=1)$ | Within truly positive cases, prediction rates should be equal across groups. | True positive parity | Qualified applicants treated unequally |
| **Calibration (CAL)** | $Y \perp\!\!\!\perp A \mid \hat{Y}$ | For each predicted score, the true positive rate should be equal across groups. | Predictive consistency | Scores misrepresent risk by group |
| **Individual Fairness (IF)** | $f(x) \approx f(x')$ if $x \approx x'$ | Similar individuals should receive similar outcomes. | Local similarity | Identical profiles, different outcomes |
| **Counterfactual Fairness (CF)** | $\hat{Y}_a(x) = \hat{Y}_{a'}(x)$ | A decision should remain the same if the individual's protected attribute were counterfactually changed. | Hypothetical independence | Decision flips if group label changes |

# Information Theory

## Linking Fairness to Quantification

- Mutual Information

  Measures how much knowing one variable reduces uncertainty about another.

  Mutual information quantifies the dependence between variables, illuminating the relationship between protected attributes and outcomes, thus serving as a foundational measure for assessing fairness.

  - If $I(A;B) = 0$: $A$ and $B$ are **independent** → knowing one tells you nothing about the other.
  - If $I(A;B)$ is large: there is **strong dependence** → one variable carries information about the other.

$$I(A;\hat{Y}) = 0 \quad \Leftrightarrow \text{perfect fairness (Demographic Parity)}$$

If $I(A;\hat{Y}) > 0$, some **information about A leaks** into the prediction — the model is partially unfair.

# Pinsker's Inequality

## Bridge between Fairness and Information Theory

Pinsker's inequality connects total variation distance to mutual information, revealing how fairness can be expressed as bounded information leakage, creating a bridge between fairness and information theory.

- Translate observable fairness gaps into information bounds.
- Express fairness as a limit on information leakage.
- Unify diverse fairness metrics under one measurable framework.

$$D_{\mathrm{TV}}(P(A,\hat{Y}), P(A)P(\hat{Y})) \leq \sqrt{\frac{1}{2} I(A;\hat{Y})}$$

Total variation distance between actual joint behavior (how predictions vary by group) and the ideal independent case (no bias) is bounded by the square root of the mutual information.

$$\text{Fairness gap}^2 \ \leq \ 2 \cdot CF \cdot I(A;Z)$$

| Fairness Metric | Traditional Definition | Unified Info-Theoretic Expression | Interpretation |
|---|---|---|---|
| **Demographic Parity (DP)** | $\hat{Y} \perp A$ | $I(A;\hat{Y}) \leq \varepsilon_{DP}$ | No information about $A$ in predictions |
| **Equal Opportunity (EO)** | $\hat{Y} \perp A \mid (Y=1)$ | $I(A;\hat{Y} \mid Y=1) \leq \varepsilon_{EO}$ | No info leakage within qualified group |
| **Calibration (CAL)** | $Y \perp A \mid \hat{Y}$ | $I(A;Y \mid \hat{Y}) \leq \varepsilon_{CAL}$ | Predictions mean same thing across groups |
| **Individual Fairness (IF)** | $f(x) \approx f(x\prime)$ | $I(A;\hat{Y} \mid X \approx X\prime) \leq \varepsilon_{IF}$ | Locally invariant outcomes |
| **Counterfactual Fairness (CF)** | $\hat{Y}_A(x) = \hat{Y}_{A\prime}(x)$ | $I(A;\hat{Y}_{do(A)}) \leq \varepsilon_{CF}$ | Causal invariance under interventions |

# Conceptual Hierarchy

## Information Leakage

$$\boxed{I(A;\hat{Y}) \;\geq\; I(A;\hat{Y} \mid Y{=}1) \;\geq\; I(A;Y \mid \hat{Y}) \;\geq\; I(A;\hat{Y}_{do(A)})}$$

| Fairness Type | Independence Condition | Typical Leakage | Interpretation |
|---|---|---|---|
| **DP** | $\hat{Y} \perp A$ | Highest $I(A;\hat{Y})$ | Weakest fairness |
| **EO** | $\hat{Y} \perp A \mid Y{=}1$ | Smaller $I(A;\hat{Y} \mid Y{=}1)$ | Stronger constraint |
| **CAL** | $Y \perp A \mid \hat{Y}$ | Smaller still $I(A;Y \mid \hat{Y})$ | Fair interpretation |
| **CF** | $\hat{Y}_A(x) = \hat{Y}_{A'}(x)$ | Ideally zero $I(A;\hat{Y}_{do(A)})$ | Strongest fairness (causal) |

# Fairness Complexity

## Understanding the Interrelationships of the Criteria

- Fairness Criteria

  The hierarchy of fairness criteria illustrates how different metrics relate to each other, showing that achieving one criterion may compromise others, emphasizing the need for a balanced approach.

- Information-Theoretic Cost

  Each fairness criterion incurs an information-theoretic cost, which affects the feasibility of achieving multiple fairness objectives simultaneously, necessitating careful consideration of trade-offs in algorithmic design.

# Impossibility Theorem

## Understanding Information Budget Constraints

- Trade-offs in Fairness

  The impossibility theorem quantifies constraints in achieving fairness, illustrating how combined fairness criteria cannot exceed total information budgets, emphasizing the inherent trade-offs in fairness metrics.

When base rates differ across groups ($I(A;Y) > 0$):

You **cannot** satisfy all fairness notions simultaneously unless the model is perfect.

$$I(A;\hat{Y}) = I(A;Y) + I(A;\hat{Y}|Y) - I(A;Y|\hat{Y})$$

- $I(A;Y)$: inherent data bias
- $I(A;\hat{Y}|Y)$: violation of EO
- $I(A;Y|\hat{Y})$: violation of calibration

# Constructive Paths

## Algorithmic Approaches to Fairness

- Representation Learning

  Leveraging information-theoretic measures, representation learning ensures that fairness constraints are met while minimizing bias in downstream model predictions, enhancing overall model reliability and fairness.

- Guaranteed Bounds

  By establishing clear limits through $I(A;Z) \leq \varepsilon^2/(2CF)$, we can ensure that models achieve fair outcomes, thereby facilitating compliance with regulatory standards and promoting ethical AI practices.

# The ε-Fair Framework

## Introducing Measurable Fairness Budget

- Information as Budget

  Fairness can be viewed as an information budget, providing a measurable framework for regulatory compliance while ensuring accountability, transparency and effective governance across algorithms and systems.

- Fairness Budget Defined

  The ε-Fair framework quantifies fairness as a bounded information leakage, allowing organizations to establish measurable budgets for various fairness types, enhancing their ability to implement fair algorithms effectively.

# Example of “Rule 80”

## Connecting practice to Pinsker’s Inequality

A protected group’s selection rate must be ≥ 80% of the reference group’s rate.

$$\frac{P(\hat{Y}=1|A=\text{minority})}{P(\hat{Y}=1|A=\text{majority})} \geq 0.8$$

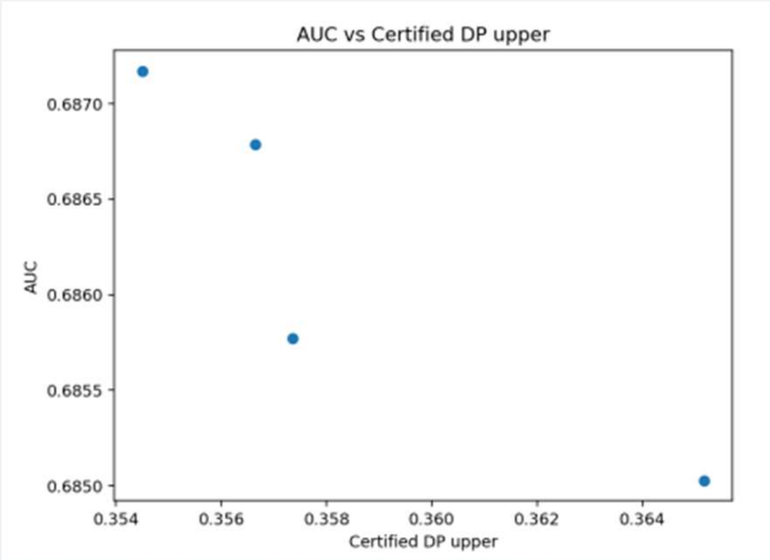
Equivalent Demographic Parity gap:

$$\Delta_{DP} = |P(\hat{Y}=1|A=0) - P(\hat{Y}=1|A=1)| \leq 0.20$$

Information-Theoretic Bridge (Pinsker’s Inequality)

$$\Delta_{DP} \leq \sqrt{2\,CF\,I(A;Z)}$$

$$I(A;Z) \leq \frac{\varepsilon^2}{2\,CF}$$




| Fairness Ratio | Max DP Gap (ε) | $I(A;Z)$ ≤ $\varepsilon^2$ / (2 CF) (for CF = 1) |
|---|---|---|
| 80 % rule | 0.20 | **0.02** |
| 90 % rule | 0.10 | 0.005 |
| 95 % rule | 0.05 | 0.00125 |

# Thank you