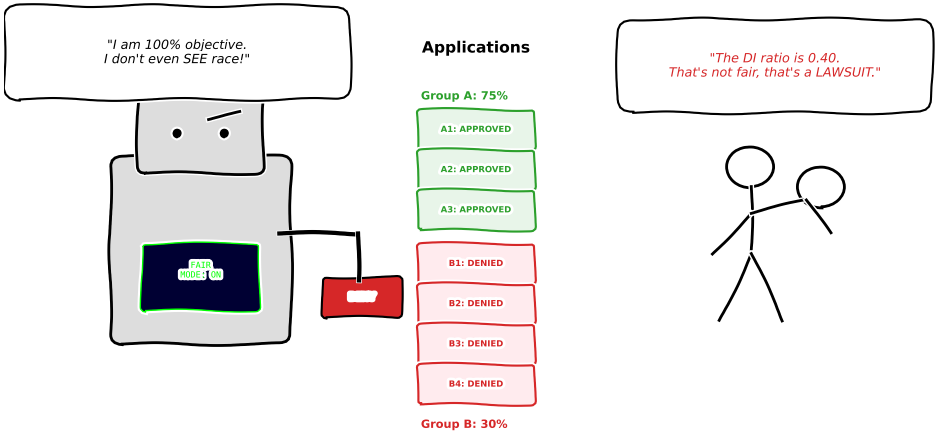
# Lesson 2.4: Algorithmic Fairness and Bias

## Module 2: The Access Problem

Prof. Dr. Joerg Osterrieder

# The Bias We Cannot See




**Removing race from the input does not remove racism from the output.**

**Algorithms inherit the biases of the data they are trained on—and sometimes amplify them.**

# Learning Objectives

By the end of this lesson, you will be able to:

1. **Identify** sources of algorithmic bias in financial decision systems
2. **Explain** how proxy variables encode protected attributes and produce disparate impact
3. **Compare** fairness metrics: demographic parity, equalized odds, and calibration
4. **Apply** explainability tools (SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME)) to audit models for hidden bias
5. **Evaluate** the fairness–accuracy tradeoff and navigate the impossibility theorem
6. **Describe** regulatory frameworks (EU AI Act, ECOA, Fair Lending) governing algorithmic decisions

**Core theme:** We have democratized access to financial services. But are the machines treating everyone fairly?

---

**These objectives span technical (metrics, tools), ethical (tradeoffs), and regulatory (compliance) dimensions.**

# From Democratization to Discrimination

**Where we are:** Technology has democratized access to financial services—mobile wallets, neobanks, and open APIs reach billions of previously unbanked people.

**The new question:** Are the algorithms behind these services treating everyone fairly?

**The Promise**

- Algorithms remove human prejudice
- Decisions are consistent and scalable
- Data-driven means objective
- Faster, cheaper underwriting

**The Reality**

- Historical data encodes historical bias
- "Objective" features act as proxies for race
- Consistency can mean consistently unfair
- Speed amplifies discrimination at scale

**The gap between promise and reality is what algorithmic fairness addresses.**

**A 2021 study found that Black and Hispanic borrowers were 40–80% more likely to be denied mortgage applications by algorithmic systems, even after controlling for creditworthiness (Bartlett et al., 2022).**

# What is Algorithmic Bias?

**Definition:** Systematic and repeatable errors in a computer system that create unfair outcomes, such as privileging one group over another.

# Historical Bias in Financial Data

**Key insight:** Models trained on historical data learn historical discrimination.

**Examples of Historical Bias**

- **Redlining (1930s–1970s):** Banks systematically denied mortgages in minority neighborhoods. Zip code data still carries this signal.
- **Credit invisibles:** 45 million Americans lack credit history. Disproportionately minority, young, and immigrant populations.
- **Income gaps:** Historical wage discrimination means lower incomes for women and minorities—models that use income inherit this.

**How Models Amplify It**

- A credit model trained on 20 years of approvals/denials learns *who was historically approved*—not who *should have been* approved
- Feedback loops: denied applicants never get loans, so no positive outcome data exists for them
- Feature selection: "years at current address" penalizes renters (disproportionately minority)
- Label bias: "default" definition may vary by product type, correlating with demographics

**The model is not biased because it is malicious. It is biased because the world that generated its training data was biased.**

# Protected Attributes and Proxy Variables

**Protected attributes** are characteristics that cannot legally be used in certain decisions.

**Protected Attributes (US/EU)**

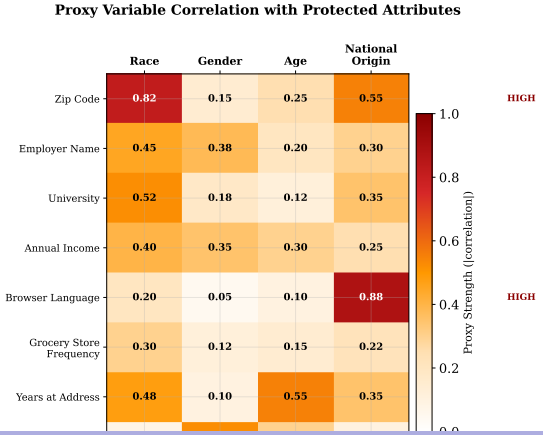
- Race / ethnicity
- Gender / sex
- Age
- Religion
- National origin
- Disability status
- Marital status (in lending)

**Key legal frameworks:**

- Equal Credit Opportunity Act (ECOA)
- Fair Housing Act (FHA)
- EU Anti-Discrimination Directives

**The Proxy Problem**

Even when protected attributes are excluded, other features can act as *proxies*:



# Disparate Treatment vs. Disparate Impact

**Two legal standards for discrimination in US law:**

**Disparate Treatment**

- **Definition:** Intentionally treating people differently based on a protected attribute
- **Example:** Using gender as a model input to price insurance differently
- **Test:** Is the protected attribute used directly?
- **Legal standard:** Almost always illegal
- **In ML:** Easy to detect—check if protected attributes are model inputs

**Disparate Impact**

- **Definition:** A facially neutral practice that disproportionately harms a protected group
- **Example:** Using zip code for credit scoring, which correlates with race
- **Test:** 80% rule (four-fifths rule)
- **Legal standard:** Illegal unless justified by business necessity
- **In ML:** Hard to detect—requires outcome analysis across groups

**The four-fifths rule:** If the selection rate for a protected group is less than 80% of the rate for the most-selected group, there is evidence of disparate impact.

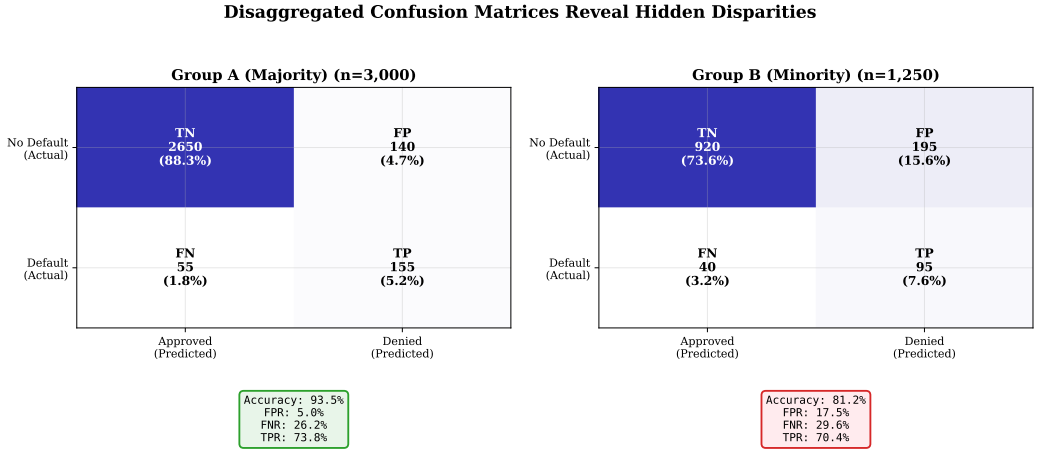
$$\text{Disparate Impact Ratio} = \frac{\text{Selection rate (disadvantaged group)}}{\text{Selection rate (advantaged group)}} \geq 0.80$$

**Example:** If Group A approval rate $= 80\%$ and Group B $= 60\%$, DI $= 60/80 = 0.75$. Since $0.75 < 0.80$, this triggers the four-fifths rule and warrants investigation.

**Most algorithmic bias cases involve disparate impact, not disparate treatment. The model does not "intend" to discriminate—but the outcomes are**

# Fairness Starts with Disaggregated Metrics

**Key insight:** Overall model accuracy can mask disparities across groups.



# Three Fundamental Fairness Metrics — Definitions

**1. Demographic Parity**

- Equal approval rates across groups
- $P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$
- Ignores qualifications entirely
- Focus: *equality of outcomes*

**2. Equalized Odds**

- Equal TPR *and* FPR across groups
- $P(\hat{Y} = 1|Y = 1, A = 0) = P(\hat{Y} = 1|Y = 1, A = 1)$
- Conditions on true label
- Focus: *equal accuracy*

**3. Calibration**

- Predicted probabilities match observed rates per group
- "70% predicted risk" means 70% default rate for all groups
- Focus: *equal meaning of scores*

**These three metrics capture different intuitions about what "fair" means. Each is mathematically precise but philosophically distinct.**

# Three Fundamental Fairness Metrics — Which Is "Right"?

**Each metric embodies a different philosophical lens:**

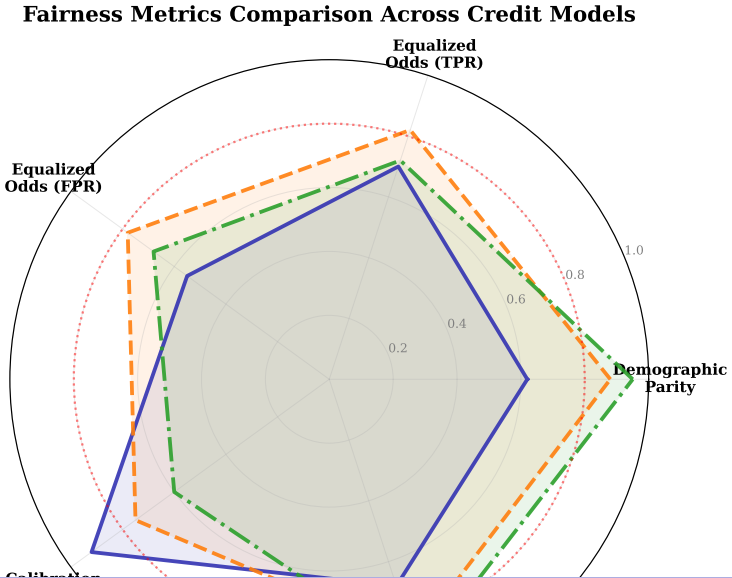
| Metric | Lens | Prioritizes |
|---|---|---|
| Demographic parity | Affirmative action | Equal outcomes regardless of qualification |
| Equalized odds | Meritocratic | Equal error rates given true outcome |
| Calibration | Actuarial | Scores mean the same thing for everyone |

**The bottom line:** The choice of fairness metric is a *value judgment*, not a technical question. Different stakeholders (regulators, lenders, advocacy groups) may reasonably prefer different metrics.

*There is no single "correct" fairness metric. The choice depends on context, stakeholders, and societal values.*

**In the next slide we will see why this choice is not just difficult but mathematically constrained: the impossibility theorem proves you cannot satisfy all three simultaneously.**

# Comparing Fairness Metrics Across Models



# The Impossibility Theorem

**Chouldechova (2017) and Kleinberg, Mullainathan, Raghavan (2016)** proved:

> **It is mathematically impossible to simultaneously satisfy demographic parity, equalized odds, and calibration—except when base rates are equal across groups.**

**Why this matters in finance:**

- Default rates differ across demographic groups (due to historical inequality)
- If base rates differ, you **must** choose which fairness criterion to prioritize
- Satisfying one metric necessarily violates another

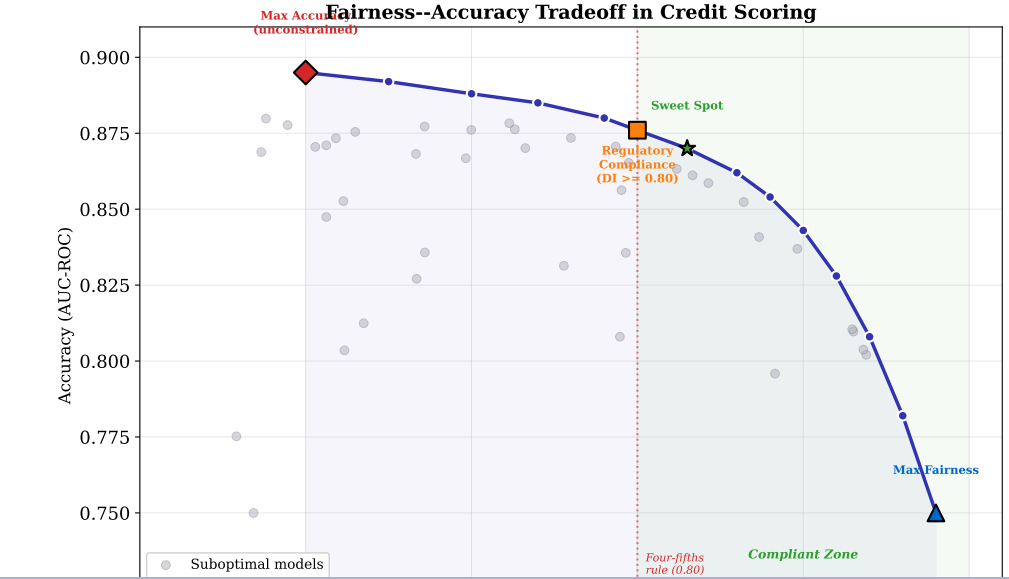
**Practical implications:**

- **Calibration + Equalized Odds** $\Rightarrow$ violates Demographic Parity (unequal approval rates)
- **Demographic Parity + Calibration** $\Rightarrow$ violates Equalized Odds (unequal error rates)
- **Equalized Odds + Demographic Parity** $\Rightarrow$ violates Calibration (scores mean different things)

---

**This is not a limitation of current technology—it is a fundamental mathematical constraint. Fairness requires explicit value choices.**

# The Fairness–Accuracy Tradeoff



# Explainability: Opening the Black Box

**Why explainability matters for fairness:**

- **Detection:** Identify which features drive disparate outcomes
- **Compliance:** ECOA requires "adverse action notices" explaining loan denials
- **Trust:** Regulators, auditors, and customers need to understand decisions
- **Debugging:** Find proxy variables hiding protected attribute signals

**Two dominant approaches:**

**SHAP (SHapley Additive exPlanations)**

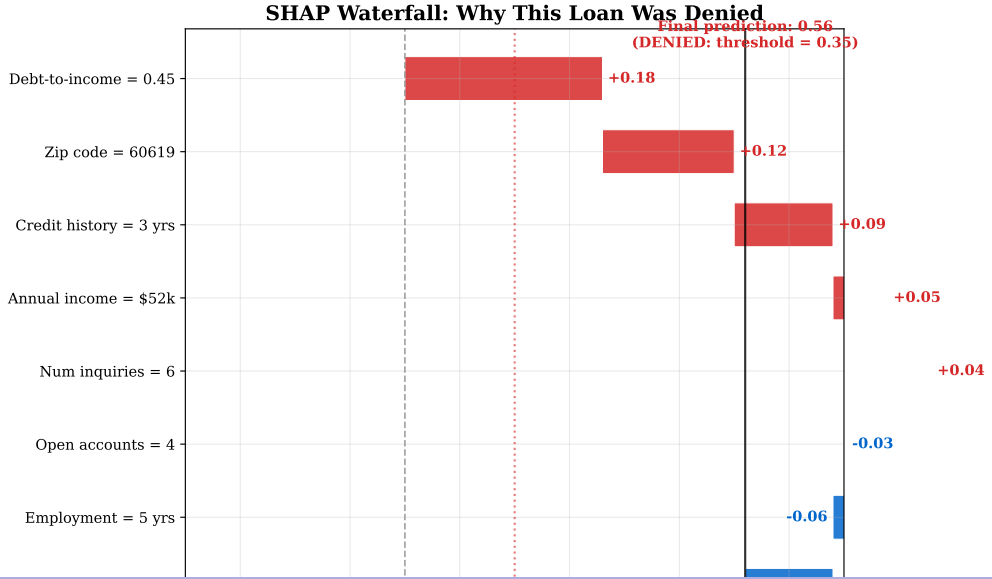
- Based on cooperative game theory (Shapley values)
- Exact: each feature's contribution to the prediction
- Model-agnostic (works with any model)
- Computationally expensive for large models
- **Global + local** explanations

**LIME (Local Interpretable Model-agnostic Explanations)**

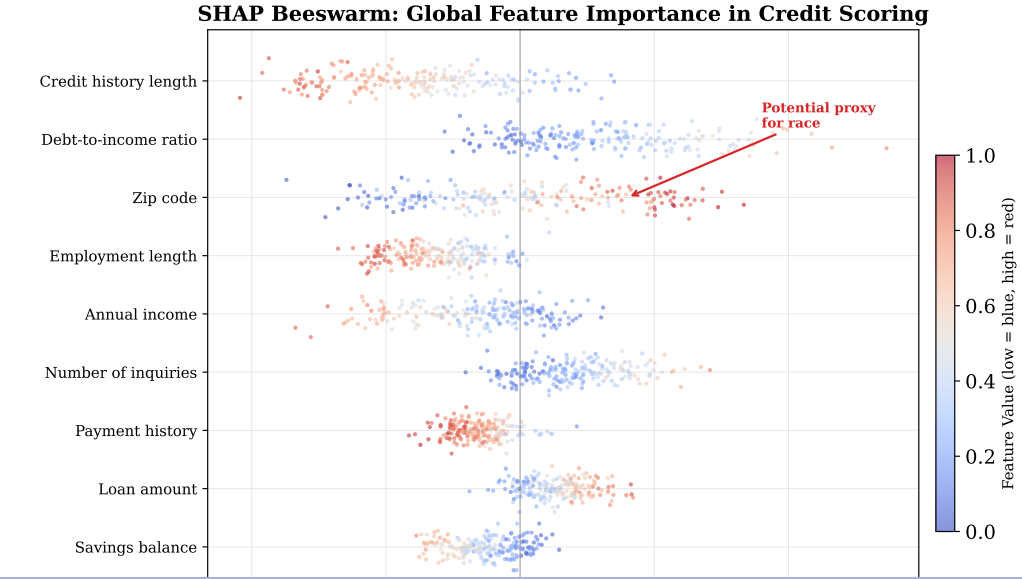
- Approximates the model locally with a simple model
- Perturbs inputs and observes output changes
- Fast and intuitive
- Explanations can be unstable (different runs $\rightarrow$ different results)
- **Local only** explanations

**ECOA requires lenders to provide specific reasons for credit denials. SHAP and LIME produce the kind of feature-level explanations that satisfy this requirement.**

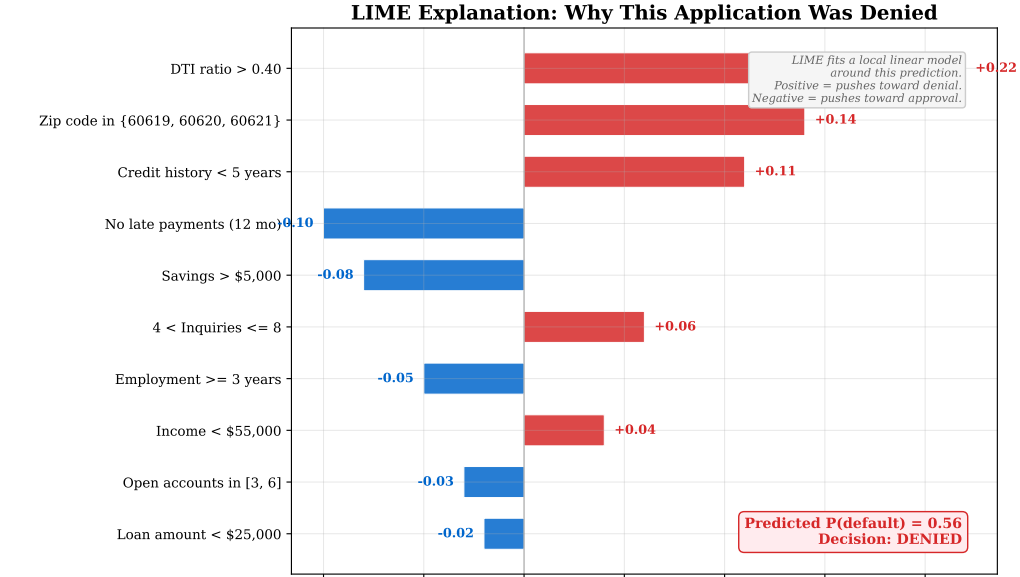
# SHAP Waterfall: Explaining a Single Decision



# SHAP Beeswarm: Global Feature Importance



# LIME: Local Explanations for Individual Decisions



# Counterfactual Fairness

**Definition (Kusner et al., 2017):** A decision is counterfactually fair if it would have been the same had the individual belonged to a different demographic group, all else being equal.

**Formal statement:**

$$P(\hat{Y}_{A \leftarrow a} = y \mid X = x, A = a) = P(\hat{Y}_{A \leftarrow a'} = y \mid X = x, A = a)$$

*"Would this person have been approved if their race/gender were different?"*

**How to test counterfactual fairness:**

- Flip the protected attribute (e.g., male → female)
- Adjust causally downstream features (e.g., income, occupation)
- Check if the decision changes
- Requires a **causal model** of how protected attributes affect other features

**Challenges:**

- Requires specifying a causal graph (contentious)
- Which features are "caused by" the protected attribute?
- Is income a "legitimate" feature or a "tainted" feature?
- Different causal assumptions → different fairness conclusions

**Counterfactual fairness is the gold standard conceptually, but it requires assumptions about causal structure that are often debatable.**

# Bias Mitigation — Fix the Data and the Algorithm

**Pre-processing** *(Fix the data)*

- Reweighting: assign higher weights to underrepresented groups
- Resampling: over/undersample to balance representation
- Disparate impact remover: transform features to remove correlation with protected attributes
- Relabeling: correct biased labels

**Pro:** Model-agnostic
**Con:** May discard useful signal

**In-processing** *(Fix the algorithm)*

- Fairness constraints during training (e.g., demographic parity penalty)
- Adversarial debiasing: train a second model to predict group membership—penalize the main model if it can
- Fair representation learning: learn embeddings that are independent of protected attributes

**Pro:** Jointly optimizes accuracy and fairness
**Con:** Algorithm-specific

**Pre-processing is the most common starting point because it is model-agnostic. In-processing yields better results but ties you to specific algorithms.**

# Bias Mitigation — Fix the Output

**Post-processing** *(Fix the output)*

- Threshold adjustment: set different decision thresholds per group to equalize outcomes
- Calibrated equalized odds: adjust outputs to satisfy equalized odds
- Reject option classification: defer uncertain predictions near the boundary

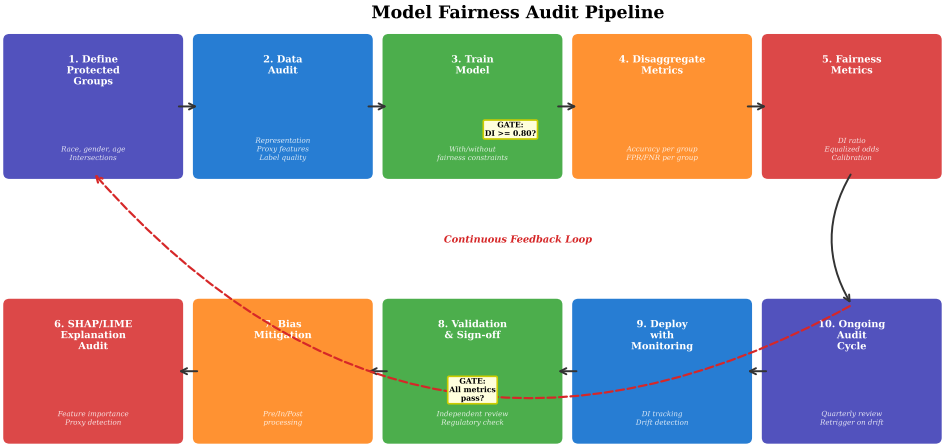
**Pro:** Simple, model-agnostic
**Con:** May violate legal standards (explicitly treating groups differently)

**Which approach to use?**

- Most production systems use a **combination** of pre-processing (data quality) and post-processing (threshold tuning)
- In-processing is strongest when you control the model architecture
- Post-processing is fastest to deploy but legally sensitive
- The right choice depends on your regulatory context

**No single strategy works universally. Most production systems use a combination of pre-processing (data quality) and post-processing (threshold tuning).**

# The Model Fairness Audit Pipeline




- **What you see:** 10-stage horizontal workflow in two rows (purple → blue → green → orange → red) from "Define Protected

# Case Study: Apple Card (2019)

**What happened:**

- Apple Card (issued by Goldman Sachs) launched in August 2019
- Tech entrepreneur David Heinemeier Hansson reported that his wife received 20x lower credit limit despite higher credit score and shared finances
- Steve Wozniak (Apple co-founder) reported the same: his wife got half his credit limit despite joint accounts

**Root cause analysis:**

- Goldman Sachs stated gender was not an input
- But proxy variables (spending patterns, income source) likely encoded gender signal
- Model was a "black box" even to Apple
- NY Department of Financial Services launched investigation

**Lessons learned:**

- "Fairness through unawareness" failed
- Explainability was absent—neither Apple nor Goldman could explain decisions
- Pre-launch fairness audit would have caught disparate impact
- Reputational cost was massive (viral Twitter thread, congressional attention)

**The NY DFS investigation concluded Goldman Sachs used a lawful algorithm but acknowledged the need for stronger disparate impact testing.**

# Case Study: COMPAS Recidivism Scores

**ProPublica Investigation (2016):**

- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) predicted recidivism risk for criminal defendants
- ProPublica found: Black defendants were **2x more likely** to be falsely flagged as high-risk (false positive rate: 44.9% vs. 23.5%)
- White defendants were **2x more likely** to be falsely labeled low-risk (false negative rate: 47.7% vs. 28.0%)

**The impossibility theorem in action:**

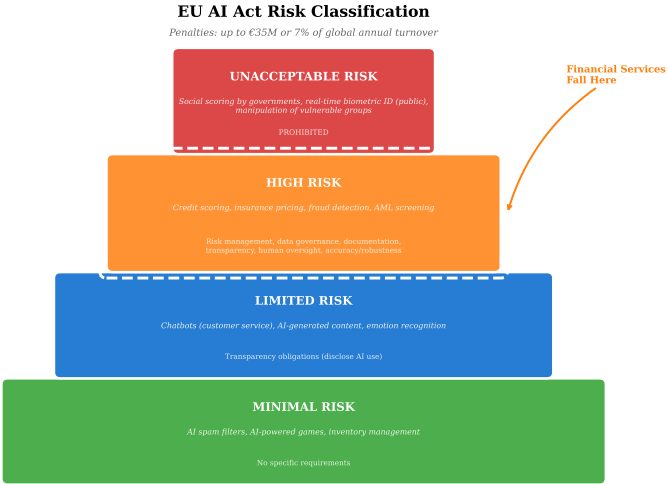
- Northpointe (COMPAS developer) showed the scores *were* calibrated: a score of "7" meant the same recidivism probability regardless of race
- ProPublica showed the scores *violated* equalized odds: error rates differed dramatically by race
- **Both sides were right.** The impossibility theorem proves these metrics cannot be satisfied simultaneously when base rates differ.

**Financial parallel:** A credit scoring model can be well-calibrated (scores mean the same thing for all groups) yet still produce disparate false positive/negative rates.

---

**COMPAS is the canonical case study in algorithmic fairness. It demonstrated that "fair" depends on which definition of fairness you choose.**

# The EU AI Act: Risk-Based Regulation




**Financial AI classifications under the EU AI Act**

# EU AI Act: What High-Risk Systems Must Do

**Six mandatory requirements for high-risk AI (including credit scoring):**

1. **Risk Management System:** Continuous identification, estimation, and mitigation of risks, including bias and fairness risks
2. **Data Governance:** Training data must be relevant, representative, and free from errors. Must test for potential biases
3. **Technical Documentation:** Detailed records of model design, training data, testing methodology, and performance metrics
4. **Transparency:** Clear information to deployers about capabilities, limitations, and intended use. End users must know they are interacting with AI
5. **Human Oversight:** Meaningful human control over the system, including the ability to override or stop the AI
6. **Accuracy, Robustness, Cybersecurity:** Performance metrics must be appropriate, consistent, and resilient to adversarial attacks

**Non-compliance penalties:** Up to €35 million or 7% of annual global turnover.

---

**The EU AI Act effectively mandates the fairness audit pipeline we described earlier. Compliance requires technical and organizational measures.**

# US Fair Lending Laws for Algorithmic Fairness

**Anti-Discrimination Laws**

- **ECOA (1974):** Prohibits discrimination in credit on the basis of race, color, religion, national origin, sex, marital status, age
- **Fair Housing Act (1968):** Prohibits discrimination in mortgage lending
- **Fair Credit Reporting Act:** Right to know why credit was denied

**Model Governance and Guidance**

- **SR 11-7:** Model risk management (includes fairness validation)
- **Consumer Financial Protection Bureau (CFPB) Guidance (2022):** Adverse action notices must be specific even for complex models

**US fair lending law is outcome-based: a facially neutral algorithm that produces disparate impact is unlawful unless justified by business necessity.**

# US Regulatory Enforcement and Trends

**Recent Enforcement Actions**

- **Upstart (2023):** CFPB scrutiny over AI lending model's disparate impact
- **Apple Card (2019):** NY DFS investigation into gender-based credit limits
- **HUD v. Facebook (2019):** DOJ sued over discriminatory ad targeting in housing
- **CFPB circular (2023):** Using complex algorithms does *not* excuse vague adverse action notices

**Emerging Trends**

- Regulators increasingly hold firms responsible for the *outcomes* of their algorithms, not just the *inputs*
- Adverse action notices must identify specific features—"the model said no" is insufficient
- State-level AI laws (e.g., Colorado AI Act) add additional requirements

**The US takes an outcome-based approach (disparate impact) rather than the EU's process-based approach (risk management). Both require fairness testing.**

# Fairness Checklist — Before and During Training

**Before Training**

- ☐ Define protected groups and fairness metric
- ☐ Audit training data for representation gaps
- ☐ Check features for proxy correlations
- ☐ Document intended use and limitations
- ☐ Establish fairness thresholds (e.g., DI $\geq 0.80$)

**During Training**

- ☐ Evaluate on disaggregated test sets
- ☐ Compare multiple fairness metrics
- ☐ Test bias mitigation strategies
- ☐ Document fairness–accuracy tradeoff

**The pre-training phase is where the most impactful fairness interventions occur. "Garbage in, garbage out" applies doubly to fairness.**

# Fairness Checklist — After Deployment and Documentation

**After Deployment**

- ☐ Monitor disparate impact ratio monthly
- ☐ Track error rates by demographic group
- ☐ Run SHAP analysis on denied applicants
- ☐ Investigate proxy variable drift
- ☐ Conduct annual third-party fairness audit

**Documentation**

- ☐ Model card with fairness metrics
- ☐ Adverse action reason codes mapped to features
- ☐ Fairness audit report with methodology
- ☐ Remediation plan for detected disparities

**This checklist aligns with both EU AI Act requirements and US ECOA/Fair Lending expectations. Adapt thresholds to your jurisdiction.**

# Open Research Questions

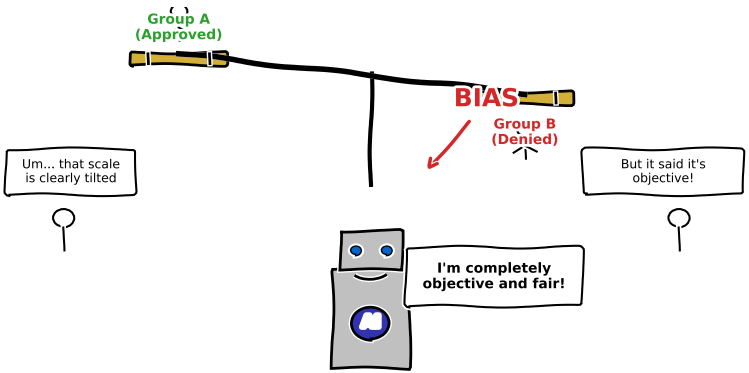
**Algorithmic fairness is an active research area. Key open problems:**

1. **Intersectionality:** How to ensure fairness for intersectional groups (e.g., Black women) when each protected attribute is fair in isolation?
2. **Long-term dynamics:** Fair decisions today may create unfair outcomes tomorrow. How to model fairness over time, accounting for feedback loops?
3. **Individual vs. group fairness:** Is it fair to treat individuals differently to achieve group-level statistics? When does positive discrimination become reverse discrimination?
4. **Causal vs. statistical fairness:** Should fairness be measured by statistical outcomes or causal counterfactuals? Different answers yield different policies.
5. **Global fairness standards:** The EU and US define fairness differently. How should multinational firms reconcile conflicting requirements?

---

**These are not purely technical questions—they require input from ethicists, legal scholars, policymakers, and affected communities.**

# One Last Thought...




**Algorithmic fairness: Just because the algorithm doesn't see bias doesn't mean it isn't there.**

**Sometimes the best way to remember a concept is to laugh about it.**

# Key Takeaways

1. **Bias enters everywhere** – from historical data, through proxy variables, to evaluation metrics. Awareness is the first step.
2. **Fairness through unawareness fails** – removing protected attributes does not remove bias. Proxy variables (zip code, employer, university) carry the same signal.
3. **Three metrics, one impossibility** – demographic parity, equalized odds, and calibration cannot be simultaneously satisfied when base rates differ.
4. **Explainability enables auditing** – SHAP waterfall plots identify which features drive individual denials. SHAP beeswarms reveal global patterns.
5. **Mitigation at three stages** – pre-processing (fix data), in-processing (constrain training), post-processing (adjust thresholds).
6. **Regulation is converging** – the EU AI Act classifies credit scoring as high-risk. US regulators focus on disparate impact outcomes.
7. **Fairness is a value choice** – technology can measure and enforce fairness criteria, but choosing *which* criteria is a human decision.

**Algorithmic fairness is not optional—it is a legal, ethical, and business imperative for any financial institution using ML.**

# Lesson Summary

**What we covered:**

- Sources of algorithmic bias: historical data, proxy variables, feedback loops, label bias
- Protected attributes, disparate treatment vs. disparate impact, and the four-fifths rule
- Three fairness metrics (demographic parity, equalized odds, calibration) and why they conflict
- Explainability tools (SHAP, LIME) for auditing model decisions
- Counterfactual fairness: "would the decision change if the person's group changed?"
- Bias mitigation strategies: pre-processing, in-processing, post-processing
- Regulatory frameworks: EU AI Act (risk-based) and US Fair Lending (outcome-based)
- Real-world failures: Apple Card, COMPAS

**The bottom line:**

*Democratizing access to financial services means nothing if the algorithms behind them perpetuate the very inequalities they were supposed to eliminate. Fairness requires intentional design, continuous monitoring, and difficult value choices.*

---

**Next lesson: we will explore data privacy, consent, and the tension between personalization and surveillance in digital finance.**