**Algorithmic Bias**

**Algorithmic bias** refers to systematic and repeatable errors in a computer system that create unfair outcomes, such as favouring one group over others. These biases can come from biased training data, flawed assumptions in the model, or imbalanced design choices.

**Real-World Examples**

**Example 1: Facial Recognition Bias**  
 Some facial recognition systems have been shown to perform poorly on people with darker skin tones. A 2018 study by MIT found that error rates for identifying Black women were as high as 34%, while white men had an error rate of less than 1%. This bias can lead to wrongful identification or exclusion in security systems.

**Example 2: Credit Scoring Systems**  
 Some credit scoring algorithms used by financial institutions have shown bias against minority communities. For example, a 2019 study showed that Black and Hispanic borrowers in the U.S. were more likely to be offered worse loan terms or denied credit altogether, even when they had similar credit profiles as white borrowers. This can happen if the AI learns from historical lending data that already reflects systemic inequality.

### Transparency vs Explainability

**Transparency** in AI means being open about how the system was built — including what data was used, how the model works, and the logic behind its decisions. It involves sharing details like algorithms, sources of training data, and who created the system.

**Explainability**, on the other hand, refers to how well humans can understand the reasoning behind an AI’s specific decision or prediction. Even if the model is complex (like deep learning), explainability focuses on making its output understandable to users.

**Why Both Matter:**

* **Transparency** builds **trust**. If developers, users, or regulators don’t know how an AI works or what data it used, it’s hard to trust the outcomes.
* **Explainability** ensures **accountability**. If an AI system rejects a loan or diagnoses a disease, people need to understand why — especially if the decision affects their lives.
* Together, they make AI systems more **ethical, fair, and responsible**, helping reduce bias and prevent harm.

**How GDPR impacts AI development in the EU**  
The General Data Protection Regulation (GDPR) shapes AI development in the EU by enforcing strict rules around data privacy and user rights.

**Key impacts include:**

* Limiting data collection to what users have explicitly consented to.
* Giving users the right to explanation for automated decisions.
* Requiring companies to document how data is processed (transparency).
* Promoting privacy by design and data minimization.

These rules encourage developers to build AI systems that respect ethics, user privacy, and fairness.

**Ethical Principles Matching**

**A) Justice**  
Fair treatment for all individuals, ensuring AI does not discriminate against any group.  
*Example: Ensuring facial recognition performs equally across all skin tones.*

**B) Non-maleficence**  
Avoiding harm by preventing misuse or dangerous consequences of AI.  
*Example: Carefully testing AI in healthcare to avoid misdiagnosis.*

**C) Autonomy**  
Respecting individuals’ rights to make informed decisions.  
*Example: Allowing users to control how their data is used.*

**D) Sustainability**  
Supporting long-term environmental, social, and economic well-being.  
*Example: Using energy-efficient AI in smart cities to reduce carbon footprint.*

**COMPAS Racial Bias Analysis Report**

**Executive Summary**

This analysis examines racial bias in the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) recidivism prediction algorithm using IBM's AI Fairness 360 toolkit. The study replicates and extends ProPublica's groundbreaking 2016 investigation, revealing significant disparities in how the algorithm treats African-American and Caucasian defendants.

**Key Findings**

Our analysis confirms substantial racial bias in COMPAS risk scoring:

* *False Positive Rate Disparity*: African-American defendants who did not recidivate were classified as high-risk at nearly twice the rate of Caucasian defendants (45% vs. 23%).
* *False Negative Rate Disparity*: Caucasian defendants who did recidivate were incorrectly classified as low-risk 70% more often than African-American defendants (48% vs. 28%).
* *Statistical Parity Difference*: -0.174, indicating significant unfairness in outcome distribution between racial groups.
* *Disparate Impact*: 0.637, well below the 0.8 threshold typically used to identify discriminatory practices.

**Predictive Performance**

The COMPAS algorithm demonstrates moderate predictive accuracy but with concerning racial disparities:

* *Overall Accuracy*: 61% for general recidivism prediction.
* *Concordance Score*: 63.6% (ability to correctly rank relative risk).
* *Racial Consistency*: While overall accuracy is similar across races, error patterns differ significantly.

**Distribution Analysis**

Risk score distributions reveal systematic bias:

* African-American defendants receive higher risk scores across all categories.
* Score distributions are more heavily skewed toward high-risk categories for African-American defendants.
* These disparities persist even when controlling for criminal history, age, and other factors.

**Remediation Recommendations**

**Immediate Actions**

1. *Algorithmic Auditing*: Implement regular bias testing using fairness metrics before deployment.
2. *Threshold Adjustment*: Calibrate decision thresholds separately for different demographic groups.
3. *Human Oversight*: Require judicial review of high-risk classifications, especially for minority defendants.

**Medium-Term Improvements**

1. *Data Collection*: Expand training data to include more diverse and representative samples.
2. *Feature Engineering*: Remove or modify features that may serve as proxies for race.
3. *Fairness Constraints*: Implement in-processing techniques like adversarial debiasing during model training.

**Long-Term Systemic Changes**

1. *Policy Reform*: Develop guidelines limiting the use of algorithmic risk assessments in sentencing decisions.
2. *Transparency Requirements*: Mandate public disclosure of algorithm performance across demographic groups.
3. *Alternative Approaches*: Explore ensemble methods that explicitly optimize for fairness alongside accuracy.

**Implementation Strategy**

* *Phase 1 (0-6 months):* Implement bias monitoring and threshold adjustments.
* *Phase 2 (6-12 months):* Deploy preprocessing techniques like reweighing.
* *Phase 3 (12-24 months):* Develop new models with fairness constraints built-in.

**Conclusion**

The *COMPAS* algorithm exhibits significant racial bias, systematically disadvantaging African-American defendants through higher false positive rates and lower false negative rates compared to Caucasian defendants. While the algorithm shows moderate predictive accuracy, these disparities raise serious concerns about fairness and equal treatment under the law. The bias mitigation techniques tested show promise in reducing these disparities while maintaining reasonable predictive performance.

**Biased Hiring Tool (Amazon)**

**Case Overview: Amazon’s Biased Recruiting AI**

In 2018, Amazon discontinued its experimental AI hiring tool after discovering that it exhibited bias against women. The system was trained on 10 years of historical hiring data, most of which reflected the male-dominated nature of the tech industry. As a result, the AI system penalized resumes that included the word "women’s" (e.g., "women’s chess club captain") and downgraded applicants from all-women colleges.

**Source of Bias**

The primary source of bias was **skewed training data**. Since the training dataset reflected past hiring decisions that were biased (favouring male applicants), the model learned to replicate and amplify those patterns. It failed to recognize qualified female candidates as equally favourable due to historical underrepresentation and gendered language in resumes.

**Three Fixes to Reduce Bias**

1. **Balanced Training Data**  
   Rebuild the dataset to ensure gender balance among successful and unsuccessful candidates. This helps the model learn to recognize qualifications across gender lines without reinforcing historical discrimination.
2. **Use of Fairness-Aware Algorithms**  
   Implement models that include fairness constraints (e.g., reject option classification or adversarial debiasing) to prevent discrimination against protected groups like gender, race, or age.
3. **Bias Auditing & Feature Sensitivity Control**  
   Exclude or modify features that act as proxies for gender, such as clubs, pronouns, or college names. Use feature attribution tools (e.g., SHAP or LIME) to understand how sensitive the model is to gendered inputs.

**Fairness Metrics to Evaluate Post-Correction**

After mitigation strategies are applied, it's essential to evaluate the model using **fairness metrics**, such as:

* **Disparate Impact Ratio (DIR)**  
  Measures whether selection rates across groups (e.g., male vs. female) are significantly different. A ratio close to 1.0 indicates fairness.
* **Equal Opportunity Difference**  
  Compares true positive rates between groups (e.g., whether qualified women and men are equally likely to be selected).
* **Demographic Parity Difference**  
  Checks if positive classification outcomes are evenly distributed across genders, regardless of qualifications.

**Conclusion**

Amazon’s case highlights how biased training data can lead to real-world discrimination when AI is used in hiring. Implementing data balancing, fairness-aware modelling, and robust fairness evaluation is critical to building equitable AI systems.