Course: AI For Software Engineering.

Group 71

Monday, July 14, 2025.

Week 7: Fair AI Systems.

Contents

AI Ethics Report

Short Answer Questions

Q1: Define algorithmic bias and provide two examples of how it manifests in AI systems.
Answer:
Algorithmic bias refers to systematic and repeatable errors in an AI system that create unfair outcomes, such as privileging one group over others. These biases often arise from biased training data, flawed assumptions in model design, or unequal data representation.
Examples:

1. **Facial recognition software** misidentifying people of color at higher rates than white individuals.
2. **Hiring algorithms** that favor male candidates over females due to historical data that reflects gender imbalance in the workforce.

Q2: Explain the difference between transparency and explainability in AI. Why are both important?
Answer:

- **Transparency** refers to the openness about how an AI system is designed, trained, and how it functions. It involves disclosing the data sources, algorithms used, and decision-making processes.
- **Explainability** is the ability to clearly describe how an AI system arrived at a specific decision or output, often in human-understandable terms.

Importance:

- **Transparency** builds trust and accountability, especially for regulators and stakeholders.
- **Explainability** ensures that users and developers can interpret and challenge AI decisions, which is critical for fairness, debugging, and compliance.

Q3: How does GDPR (General Data Protection Regulation) impact AI development in the EU?
Answer:
GDPR affects AI by enforcing strict data protection and privacy rules. AI systems must comply with requirements such as:

- **Data minimization** (using only necessary data)
- **User consent** for data collection and processing

- **The right to explanation**, where individuals can ask for clarification on automated decisions affecting them
  These rules encourage ethical AI development but also limit certain data-driven innovations unless privacy safeguards are in place.

Ethical Principles Matching

Principle	Definition
A) Justice	Fair distribution of AI benefits and risks.
B) Non-maleficence	Ensuring AI does not harm individuals or society.
C) Autonomy	Respecting users' right to control their data and decisions.
D) Sustainability	Designing AI to be environmentally friendly.

Case Study – Amazon's Biased AI Recruiting Tool

Scenario:
Amazon developed an AI recruitment tool that unintentionally penalized female candidates.

1. Source of Bias

- **Training Data Bias:**
  The model was trained on resumes submitted over 10 years, mostly from male applicants in tech roles. This created a gender-skewed data pattern.
- **Model Design Oversight:**
  There was no in-built mechanism to detect or correct for gender bias during model development or testing.

2. Proposed Fixes for Fairness

Fix 1: Clean and Rebalance Training Data

- Remove gender indicators (e.g., names, women's college references).
- Add resumes from a gender-diverse applicant pool to train the model fairly.

Fix 2: Integrate Fairness-Aware Algorithms

- Use fairness constraints (like equal opportunity) during model training.
- Adopt de-biasing techniques such as adversarial learning or data re-weighting.

Fix 3: Institutionalize Regular Audits

- Establish ongoing audits by cross-functional teams (AI, HR, legal).
- Test outcomes across demographics before and after deployment.

3. Fairness Evaluation Metrics

Metric	Purpose
Demographic Parity	Ensure selection rates are similar across gender groups.
Equal Opportunity	Confirm all qualified candidates have equal chances, regardless of gender.
Disparate Impact Ratio	Ratio of female to male candidate selection rates (target: 0.8–1.0).
False Positive/Negative Rates by Group	Compare errors across groups to detect bias in misclassification.

Case Study – Facial Recognition in Policing

Scenario:
A facial recognition system used in policing misidentifies individuals from minority groups at significantly higher rates.

1. Ethical Risks

Risk	Description
Wrongful Arrests	Misidentifications can lead to innocent individuals—especially minorities—being wrongly detained or arrested.
Privacy Violations	Constant surveillance may infringe on individuals' rights to privacy, especially when used without consent or warrants.
Discrimination and Bias	Disproportionate errors against minorities reinforce systemic discrimination and social injustice.

Risk	Description
Lack of Accountability	If the system makes incorrect decisions and no clear human oversight exists, there's limited recourse for affected individuals.
Erosion of Public Trust	Communities may lose trust in law enforcement and technology, reducing cooperation and amplifying tension.

Recommended Policies for Responsible Deployment

Policy 1: Mandatory Human Oversight

- Facial recognition results must **never be the sole basis** for action—officers must verify identity through additional, non-AI means.

Policy 2: Regular Bias Audits

- Conduct **independent evaluations** on system performance across demographics. If bias is found, retrain or pause deployment until fixed.

Policy 3: Transparency and Disclosure

- Clearly disclose to the public when and where facial recognition is used. Publish performance metrics and incident reports regularly.

Policy 4: Consent and Legal Safeguards

- Use facial recognition **only with legal authorization** (e.g., court warrants) or **explicit public consent** in non-criminal settings.

Policy 5: Community Involvement

- Engage with civil rights groups, ethicists, and the communities most affected in policymaking and deployment decisions.

Report: Racial Bias in COMPAS Recidivism Risk Scores

This project investigated potential racial bias in the COMPAS recidivism dataset using IBM's AI Fairness 360 (AIF360) toolkit. The dataset, originally published by ProPublica, contains information on defendants' criminal histories, demographics, and COMPAS risk assessment scores.

Racial Bias in COMPAS Recidivism Risk Scores.

After applying standard preprocessing steps — including filtering out incomplete records, limiting cases to those within 30 days of arrest, and removing minor charges — a binary race attribute was introduced: Caucasian as the privileged group and all other races as the unprivileged group. Categorical variables such as sex, age category, charge degree, and score text were encoded numerically to comply with AIF360 requirements.

Fairness was evaluated using two key metrics: **Mean Difference** and **Disparate Impact**. The analysis revealed a **Mean Difference of approximately -0.19**, indicating that non-Caucasian individuals were less likely to receive favorable outcomes. A **Disparate Impact score of around 0.76** was also recorded, falling below the commonly accepted threshold of 0.80, suggesting a meaningful disparity in treatment between racial groups.

These findings support existing concerns that algorithmic risk scoring systems like COMPAS may perpetuate racial bias, particularly against minority groups, leading to unfair criminal justice decisions.

To mitigate this bias, we recommend:

1. **Reweighing** the dataset during preprocessing to balance outcomes between privileged and unprivileged groups.
2. **Applying fairness-aware algorithms** such as adversarial debiasing or prejudice remover regularization during model training.
3. **Conducting regular audits** to ensure ongoing fairness in deployed systems.

Future work may include comparing classifier performance with and without mitigation steps, visualizing disparities, and engaging stakeholders to define acceptable fairness thresholds in high-stakes domains like criminal justice.

Policy Guideline: Ethical Use of AI in Healthcare

1. Patient Consent Protocols

- **Informed Consent**: Patients must be informed when AI is involved in diagnosis, treatment recommendations, or decision-making processes. This includes disclosing the AI system's role, capabilities, and limitations.
- **Opt-In Model**: Use an explicit opt-in approach for AI-assisted services. Patients should be given a choice to approve or decline AI involvement.

- **Data Usage Consent**: Any use of patient data to train or improve AI systems must require clear, written consent. Data should be anonymized wherever possible.

2. Bias Mitigation Strategies

- **Representative Training Data**: Ensure datasets reflect diverse populations (e.g., race, age, gender, socio-economic status) to reduce systematic bias.
- **Regular Bias Audits**: Implement routine checks to detect and address disparities in outcomes across different demographic groups.
- **Fairness-Aware Models**: Use algorithms that include fairness constraints (e.g., reweighing, equalized odds) during development and deployment.
- **Third-Party Review**: Encourage external validation of AI models to ensure objectivity and adherence to equity principles.

3. Transparency Requirements

- **Explainability**: AI systems must offer interpretable results understandable to both healthcare providers and patients. Black-box models should be avoided in critical care settings.
- **Model Disclosure**: Document the AI system's intended use, data sources, performance metrics, and known limitations. Share this information with medical staff and oversight bodies.
- **Traceability and Accountability**: Maintain detailed logs of AI decisions to enable auditing. Clearly define human oversight responsibilities for all AI-assisted decisions.

Enforcement & Oversight:
This policy should be enforced by healthcare ethics boards in partnership with technical teams. Compliance should be reviewed annually and updated as technologies evolve.

Conclusion

As AI systems become increasingly integrated into critical areas such as hiring and law enforcement, ensuring fairness, transparency, and accountability is no longer optional—it is essential. The cases analyzed reveal how algorithmic bias, if left unchecked, can lead to real-world harm, including discrimination and the erosion of public trust. Through responsible design, inclusive data practices, and strong ethical policies, we can guide AI development toward outcomes that are just, explainable, and respectful of human rights. The future of

ethical AI depends not only on technical solutions but on active human oversight, legal safeguards, and continuous dialogue with the communities it affects.