Define algorithmic bias and provide two examples of how it manifests in AI systems.

Algorithmic bias refers to systematic errors in AI algorithms that lead to unfair or discriminatory outcomes, often reflecting and amplifying existing societal prejudices. This bias can arise from unrepresentative training data, biased labeling or inappropriate model assumptions.

Examples of Algorithmic Bias.

1. **Healthcare Disparities**: An AI algorithm used for patient care may under-refer Black patients compared to white patients with similar health conditions, reflecting historical biases in the training data.
2. **Hiring Algorithms**: Amazon's AI recruiting tool was found to favor male candidates because it was trained on resumes from a predominantly male workforce, leading to gender bias in hiring decisions.

Explain the difference between transparency *and* explainability in AI. Why are both important?

Transparency and explainability are essential concepts in AI, each serving a unique purpose:

Transparency refers to the openness about how an AI system operates, including its design, data sources, and decision-making processes.

It builds trust among users by providing clear information about how AI systems function, which is crucial for ethical compliance and accountability.

Explainability focuses on providing understandable reasons for specific decisions made by an AI system, clarifying the logic behind individual outcomes.

It ensures that users can comprehend and trust the decisions made by AI, which is particularly vital in high-stakes areas like healthcare and finance.

They both matter in that Transparency fosters overall trust in AI systems, while explainability enhances trust in specific outputs. Together, they ensure ethical use and compliance with regulations, ultimately leading to more reliable AI applications.

How does GDPR (General Data Protection Regulation) impact AI development in the EU?

- **Data Protection by Design**:GDPR requires that data protection measures be integrated into AI systems from the design phase, ensuring privacy is prioritized throughout the AI lifecycle.

- **Consent and Legal Grounds**:Organizations must obtain explicit consent from individuals for processing personal data. The GDPR outlines specific legal bases for data processing, which must be adhered to during AI development.

- **Data Minimization**:AI systems are mandated to collect only the minimum necessary personal data for their intended purposes, which can be challenging given the data-intensive nature of AI.

- **Individual Rights:** GDPR grants individuals rights over their data, including access, rectification, and the right to explanation regarding automated decisions. AI systems must be designed to respect these rights.

- **Accountability and Compliance**: Organizations must conduct Data Protection Impact Assessments (DPIAs) for high-risk AI applications and maintain documentation to demonstrate compliance with GDPR.

- **Regulatory Framework**: The GDPR sets a precedent for future AI regulations, such as the proposed EU AI Act, which aims to ensure that AI systems are safe, transparent, and respect fundamental rights.

Guideline for Ethical AI Use in Healthcare.

Patient Consent Protocols

- Informed Consent: Obtain explicit consent from patients before using AI tools, ensuring they understand how their data will be utilized and the implications for their care.
- Ongoing Consent: Establish mechanisms for continuous consent, allowing patients to withdraw their data at any time without affecting their treatment.
- Clear Communication: Use accessible language to explain AI's role in patient care, considering demographic factors that may influence understanding.

Bias Mitigation Strategies

- Diverse Data Sets: Train AI systems on inclusive datasets that represent various demographics to minimize bias and ensure equitable healthcare outcomes.

- Regular Audits: Conduct ongoing evaluations of AI systems to identify and address biases in decision-making processes.
- Stakeholder Engagement: Involve a diverse group of stakeholders, including ethicists and community representatives, in the development and assessment of AI technologies.

Transparency Requirements

- Algorithmic Transparency: Clearly disclose the algorithms used, data sources and decision-making processes of AI systems to foster trust and accountability.
- Education for Patients and Providers: Ensure both patients and healthcare providers understand AI functionalities, limitations, and potential risks.
- Public Reporting: Regularly publish performance metrics and findings related to AI systems, including any identified biases and corrective actions taken.

By adhering to these guidelines, healthcare organizations can promote ethical AI use, ensuring that it enhances patient care while safeguarding rights and promoting equity.

Matching the ethical principles with their corresponding descriptions.

1. **Justice**: Fair distribution of AI benefits and risks.

2. **Non-maleficence**: Ensuring AI does not harm individuals or society.

3. **Autonomy**: Respecting users' right to control their data and decisions.

4. **Sustainability**: Designing AI to be environmentally friendly.

Amazon's AI recruiting tool penalized female candidates.

Identify the source of bias (e.g., training data, model design).

- **Training Data**: The AI recruiting tool was trained predominantly on resumes submitted over a ten-year period, which reflected a male-dominated applicant pool. This historical bias in the training data led the algorithm to favor male candidates and systematically downgrade female candidates, particularly those whose resumes included terms like "women's"

Propose three fixes to make the tool fairer.

- **Diverse Training Data**: Ensure that the training dataset includes a balanced representation of genders and other demographics. This could involve actively sourcing resumes from underrepresented groups and ensuring that the dataset reflects a more equitable distribution of candidates across genders and backgrounds.

- **Bias Auditing and Correction Mechanisms**: Implement regular audits of the AI system to identify and correct biases. This could include using techniques such as adversarial debiasing, where the model is trained to minimize bias while maintaining performance accuracy. Additionally, employing fairness constraints during the model training process can help ensure that the outcomes are equitable.

- **Human Oversight and Decision-Making**: Introduce a layer of human review in the recruitment process, especially for candidates flagged by the AI system. Recruiters should be trained to recognize potential biases and make final decisions based on a holistic view of the candidate's qualifications, rather than solely relying on AI-generated scores.

Suggest metrics to evaluate fairness post-correction.

- **Demographic Parity**: Measure the acceptance rates of candidates across different demographic groups (e.g., gender, race) to ensure that the AI system does not favor one group over another. The goal is to achieve similar acceptance rates for all groups.

- **Equal Opportunity Metrics**: Evaluate the true positive rates for different demographic groups. This metric assesses whether qualified candidates from all backgrounds have an equal chance of being selected for interviews or job offers.

- **Disparate Impact Ratio**: Calculate the ratio of selection rates between different demographic groups. A ratio close to 1 indicates that the AI system is treating all groups fairly, while significant deviations suggest potential bias.

Ethical Risks of Misidentifying Minorities with Facial Recognition Systems

Facial recognition technology (FRT) poses several ethical risks, particularly regarding its misidentification of minorities:

- **Wrongful Arrests**: Higher misidentification rates can lead to wrongful arrests, who were falsely accused due to flawed facial recognition matches.

- **Privacy Violations**: FRT often operates without individuals' consent, infringing on privacy rights and leading to unwarranted surveillance.

- **Discrimination**: The technology tends to be less accurate for people of color, perpetuating systemic biases and eroding trust in law enforcement.

- **Chilling Effect**: The fear of being monitored can deter individuals from participating in public activities, undermining free expression..

Recommended Policies for Responsible Deployment

- **Regulatory Frameworks**: Establish clear regulations requiring warrants for FRT use in law enforcement and mandate transparency about its application.

- **Bias Mitigation**: Require regular audits of FRT systems to identify and correct biases, ensuring diverse training datasets.

- **Public Awareness**: Inform individuals when FRT is used in public spaces and require consent for data collection.

- **Limit Use Cases**: Restrict FRT deployment to serious crimes, preventing its use for minor offenses or general surveillance.