Assignment: Designing Responsible and Fair AI Systems (Group Report)

Part 1: Theoretical Understanding

1. Short Answer Questions

Q1: Define algorithmic bias and provide two examples of how it manifests in AI systems.

Algorithmic bias refers to systematic and repeatable errors in a computer system that create unfair outcomes, such as preferring one arbitrary group of users over others. This bias is not a flaw in the code execution itself, but rather a reflection of societal, historical, or data collection biases embedded in the training data, the algorithm's design, or its objective function.

Two Examples of Manifestation:

1. **Hiring Tools (Gender Bias):** An AI tool trained on decades of hiring data from a male-dominated industry (e.g., tech engineering) learns to associate words or traits common among successful male applicants with "high performance." When evaluating new candidates, it may systematically penalize résumés containing female-associated traits (like the word "women's") or filtering out female candidates, regardless of qualifications.
2. **Lending Decisions (Racial/Socioeconomic Bias):** A loan application AI is trained using data that shows lower repayment rates in certain zip codes predominantly inhabited by minority groups. The AI may learn to use zip code as a proxy for race or income and unjustly deny loans to qualified individuals from these areas, perpetuating historical financial inequities.

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

Concept	Definition	Focus
Transparency (Interpretability)	Understanding how the AI system works. This involves knowing the data, the objective function, and the algorithm used (e.g., knowing a model is a decision tree or a neural network).	The system itself and its internal mechanics.
Explainability (XAI)	Understanding why a specific decision was made for a specific input. This involves providing human-understandable justification for the output (e.g., "The model predicted low risk because the user has 10 years of credit history and a low debt-to-income ratio").	The decision and its justification.

Importance:

- **Transparency** is crucial for **accountability and auditing**. Knowing the underlying mechanics allows regulators or auditors to check for malicious intent or design flaws.

- **Explainability** is crucial for **trust, recourse, and utility**. Users need to trust the system, and those negatively affected by a decision must have the right to challenge it with a concrete reason (right to explanation).

Q3: How does GDPR (General Data Protection Regulation) impact AI development in the EU?

GDPR profoundly impacts AI development by setting strict requirements for processing personal data, which is the lifeblood of most AI systems. Key impacts include:

1. **Lawful Basis for Processing:** AI developers must have a specific, legal reason (usually explicit, informed **consent**) to use personal data for training models.
2. **Right to Be Forgotten:** Individuals have the right to request that their personal data be erased. This is technically challenging for AI, as removing a data point often means retraining the entire model.
3. **Right to Explanation/Automated Individual Decision-Making:** Article 22 grants data subjects the right not to be subject solely to automated decisions (like loan denials or hiring filters) without human intervention, and the right to understand the logic behind the decision. This drives the need for **Explainable AI (XAI)** in critical applications.
4. **Data Minimization & Pseudonymisation:** Forces developers to only collect the data strictly necessary for the purpose and to use anonymization techniques wherever possible, limiting the quality and quantity of data available for model training.

2. Ethical Principles Matching

Match the following principles to their definitions:

- **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.
- **A) Justice:** Fair distribution of AI benefits and risks.

Part 2: Case Study Analysis

Case 1: Biased Hiring Tool (Amazon)

Scenario: Amazon's AI recruiting tool penalized female candidates.

1. Identify the source of bias:

The primary source of bias was **Historical Data Bias (Selection Bias)**. The model was trained on résumés submitted over a 10-year period, predominantly sourced from a male-dominated tech industry. The AI learned that *past* successful hires were mostly male, leading it to conclude that male gender was a predictor of success. Consequently, it penalized résumés containing words more common to female applicants, like "women's" (e.g., "captain of the women's chess club"). The bias was an accurate reflection of historical inequity, but an

unacceptable design flaw for a futuristic hiring tool.

2. Propose three fixes to make the tool fairer:

1. **Pre-processing (Attribute Masking/Removal):** Completely remove or mask any sensitive or proxy features related to gender (e.g., names, gendered language, references to gender-specific colleges/groups) from the input data *before* training.
2. **In-processing (Adversarial Debiasing):** Use an adversarial neural network during training. This technique trains the main classification model simultaneously with an "adversary" model whose job is to predict the protected attribute (gender) from the output. The main model is penalized if the adversary succeeds, forcing the classifier to learn features that are independent of gender.
3. **Post-processing (Equalized Odds):** Calibrate the model's output scores post-training to ensure that the True Positive Rate (TPR) and False Positive Rate (FPR) are roughly equal for both male and female candidates. This ensures the model is equally accurate in identifying high-performers regardless of gender.

3. Suggest metrics to evaluate fairness post-correction:

1. **Disparate Impact Ratio (DIR):** This is the ratio of the selection rate for the unprivileged group (e.g., female candidates) to the selection rate for the privileged group (e.g., male candidates). A common threshold for fairness is .
2. **Equal Opportunity Difference (EOD):** Measures the difference in **True Positive Rate (TPR)** between the privileged and unprivileged groups. A fair system should have a TPR near zero, meaning it successfully selects equally high-performing candidates regardless of the protected attribute.

Case 2: Facial Recognition in Policing

Scenario: A facial recognition system misidentifies minorities at higher rates.

1. Discuss ethical risks:

1. **Wrongful Arrests and Incarceration:** If the system is trained on predominantly light-skinned faces, it performs significantly worse (higher False Positive Rates) on darker-skinned individuals. A misidentification can lead directly to the wrongful arrest, detention, and permanent psychological harm of an innocent person.
2. **Disproportionate Surveillance and Mission Creep:** The deployment of FRT often leads to over-policing and targeted surveillance of minority communities. This results in **chilling effects** on civil liberties, discouraging peaceful protest and association. Over time, the technology inevitably expands beyond its original intent (mission creep) into mass surveillance.
3. **Lack of Autonomy/Right to Recourse:** The opacity of the underlying AI model and the often irreversible nature of a police investigation based on FRT severely restricts an individual's **autonomy** and their ability to seek **recourse** or explanation for the machine-generated accusation.

2. Recommend policies for responsible deployment:

1. **Mandatory Third-Party Audits and Certification:** Before deployment, the system must undergo independent, verifiable audits to certify performance across diverse demographic groups (race, gender, age). If the Equal Opportunity Difference (EOD) or other fairness metrics fall below an acceptable threshold for any group, deployment is prohibited.
2. **Strict Human-in-the-Loop Requirement:** FRT must only be used as a *lead* generation tool, never as a definitive form of identification. A highly trained human expert must verify the AI's match using secondary, non-AI forms of evidence before any legal action is taken.
3. **Narrow Use Case Restrictions:** Deployment must be restricted to specific, high-stakes scenarios (e.g., immediately locating a missing child or a known terrorist suspect) and banned from generalized public surveillance or tracking political dissent.

Part 4: Ethical Reflection

Prompt: Reflect on a personal project (past or future). How will you ensure it adheres to ethical AI principles?

Reflection:
If I were to build a Personalized News Aggregator AI aimed at reducing filter bubbles, my primary ethical challenge would be Autonomy and Non-maleficence (avoiding manipulation). To ensure adherence to ethical principles, I would implement the following:

1. **Transparency and User Autonomy:** I would not hide the recommender system's mechanics. Instead of silently optimizing for "click-bait," the system would offer a "Why This Story?" button explaining the ranking factors (e.g., "You read a lot about space exploration, and this topic challenges your known views on climate policy"). Users would also be given explicit controls to adjust their filter bubble radius, putting **autonomy** in their hands.
2. **Bias Mitigation:** I would audit the source data (the publishers) to ensure diversity of political and geographic origin. I would use the **Group Fairness** principle to ensure that the algorithm does not disproportionately suppress content from smaller, unprivileged media sources in favor of large, dominant ones, thereby ensuring a fair distribution of visibility (**Justice**).

Bonus Task: Policy Proposal

Guideline for Ethical AI Use in Healthcare: The CARE Framework

Title: The CARE Framework: Consent, Accountability, Resolution, and Equity in Clinical AI

Goal: To govern the deployment of AI systems (diagnosis, triage, robotic surgery planning) within the healthcare sector, prioritizing patient well-being and trust.

1. Patient Consent Protocols (C: Consent and Autonomy)

- **Layered and Specific Consent:** Patient consent for data use must be explicit, informed, and separate from consent for treatment. Patients must consent specifically to the *type* of AI use (e.g., "AI for diagnostic screening" vs. "AI for treatment recommendation").
- **Right to Opt-Out:** Patients must be given the unambiguous right to opt-out of having a fully automated decision made about their care. If a decision involves AI, a qualified human clinician must review and sign off on it.
- **Data Erasure (Technological Feasibility):** Healthcare providers must document their technical capacity for the "Right to Be Forgotten." While full erasure from a model is difficult, procedures must be in place to remove the patient's records from the training dataset and future model iterations upon request.

2. Bias Mitigation Strategies (E: Equity and Justice)

- **Diverse Data Mandate:** All AI models must be trained and tested on datasets that are demographically and clinically representative of the target patient population (including age, gender, race, socioeconomic status, and co-morbidities).
- **Fairness Auditing:** Developers must use the **Equal Opportunity Difference (EOD)** as the primary fairness metric. The False Negative Rate (missing a diagnosis) for the unprivileged group (e.g., a minority or rural population) must not exceed a margin of compared to the privileged group.
- **Bias Remediation:** If significant performance disparities are detected (particularly in diagnostic sensitivity), the model must not be deployed until pre-processing (re-weighting) or post-processing (score calibration) mitigation strategies are implemented and verified by a third-party ethics board.

3. Transparency Requirements (A: Accountability and Resolution)

- **Clinical Explainability:** For all high-risk clinical decisions (e.g., cancer diagnosis, triage priority), the AI system must generate an **Explainable AI (XAI)** report outlining the top five input factors (e.g., lab values, imaging features) that contributed to the recommendation. This must be accessible to the reviewing clinician.
- **Accountability Framework:** Clear contractual responsibility must be established *before* deployment. If an AI system contributes to patient harm, accountability must be traceable to the party responsible for the system's design, training, or deployment (Developer, Hospital, or Manufacturer), ensuring the patient has a clear path for **resolution and compensation**.