Part 1: Theoretical Understanding

1. Short Answer Questions

Q1 Algorithmic Bias

Algorithmic bias is a silent distortion in AI systems where unfair preferences or prejudices seep into decisions, often mirroring historical inequalities or flawed data.

Examples

- Job Screening AIs rejecting resumes with "women's university" affiliations, perpetuating gender gaps.
- Healthcare Algorithms prioritizing white patients over sicker Black patients due to racially skewed training data.

Q2. Transparency vs. Explainability

Transparency: Pulling back the curtain on an AI's design like revealing the ingredients in a recipe. Example: Publicly sharing an AI's data sources.

Explainability: Decoding why an AI made a decision—like a doctor explaining a diagnosis in plain terms. Example: A loan-rejection AI detailing, "Your debt-to-income ratio exceeded 45%."

Why both are important: Transparency builds societal trust; explainability empowers users to challenge unfair outcomes. Without them, AI becomes an inscrutable "black box" of unchecked power.

Q3. GDPR's AI Impact in the EU

GDPR forces AI developers into a tightrope walk:

- Consent-Centric Design: AI must ask, "May I use your data?"—not assume silence means yes.
- Minimize Data Hunger: Collect only essential data (e.g. a fitness app tracking steps, not your location).
- Right to Explanation: Users can demand, "Why did your AI reject my mortgage application?"

- Bias Audits: Regular checks for discrimination or face fines up to 4% of global revenue.

2. Ethical Principles Matching

Imagine AI ethics as a compass guiding a ship through fog:

- Non-maleficence (B) acts as the hull, shielding society from harm—ensuring AI avoids toxic misinformation or unsafe decisions.
- Autonomy (C) is the helm, placing control in users' hands—like letting them edit data or opt out of profiling.
- Sustainability (D) maps the course toward green horizons—minimizing AI's carbon footprint via efficient data centres.
- Justice (A) balances the cargo, distributing AI's benefits (e.g. healthcare access) and risks (e.g. job loss) equitably across all communities.

Part 2: Case Study Analysis

Case 1: Biased Hiring Tool

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

Task 1: Identify the source of bias

The bias stemmed from three interconnected sources:

1.1 Training Data: 10 years of resumes from male-dominated tech roles taught the AI that "successful candidates" were overwhelmingly male.

1.2 Model Design: It penalized phrases associated with women (e.g. "women's chess club") and rewarded "masculine-coded" verbs (e.g. "executed").

1.3 Feedback Loop: Rejecting female candidates reduced female examples in future data, deepening the bias.

Task 2: Propose three fixes

2.1 Rebalance Training Data: Add synthetic resumes featuring non-stereotypical female career paths and partner with women-in-tech groups for real-world data.
2.2 Bias Auditing Layer: Integrate real-time checks that flag gender-correlated rejections for human review.
2.3 Redefine "Merit": Retrain the model to value collaborative skills (e.g. "mentored" "organized") over aggressive language.

Task 3: Suggest fairness metrics

- Demographic Parity: % of female candidates in "recommended" lists vs. application pool.
- False Negative Rate Gap: Difference in rejection rates for equally qualified men vs. women.
- Rescue Rate: % of AI-rejected female resumes overturned by human reviewers.

Case 2: Facial Recognition in Policing

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

Task 1: Discuss ethical risks

Three critical risks

- Wrongful Arrests: Higher false positives for minorities could lead to detainment of innocent people (e.g., Robert Williams' 2020 case in Detroit).

- Privacy Erosion: Mass scanning in public spaces (e.g. protests) creates permanent biometric databases without consent.
- Reinforced Discrimination: Over-policing minority neighbourhoods due to false matches entrenches systemic racism.

Task 2: Recommend deployment policies

- Legally Restricted Use: Only allow scans for violent felonies with judicial warrants. Ban surveillance of protests/public gatherings.
- Accuracy Equity Mandate: Decommission systems with >0.1% error rate disparities across racial groups.
- Transparency & Redress:

  Public log of all scans (location, purpose, result).

  Independent audits by civil rights groups.

  Compensation fund for victims of misidentification.

Part 4: Ethical Reflection

Part 4: Ethical Reflection - My Weather Prediction App

For my personal project an AI that predicts extreme weather for farmers I'd ensure ethics through these practical steps:

4.1 **Talk to real farmers first**

Before coding, I'd visit 10+ small farms to understand their needs. Does the app help organic rice growers in Thailand as much as Iowa corn farmers?

4.2 **Bias checkpoints**

- Test predictions for neighborhoods of different income levels
- Compare accuracy for tropical vs. temperate climates
- Hire migrant farmworkers as beta testers

4.3 **Transparency sandwich**

Every prediction shows:

- How we know: "This flood alert uses satellite + soil moisture data"
- What's iffy: "Low confidence near mountains - check local radio"
- Who to call: Local emergency contacts auto-updated.

4.4 **Safety nets**

Partner with farm co-ops to provide:

Free rain gauges for areas with spotty data

Disaster loans when predictions fail

Bonus Task: Healthcare AI Guidelines

Our Promise: AI That Cares Like Humans Do

1. Patient Consent Rules

1.1 **Plain language explanations**

- Show videos like: "This AI helps find tumors. It sees your scans but not your name."
- Use picture menus for kids/non-readers

1.2 **Re-consent every year**

Text reminder: "Our AI learned new things. Still OK to use your data?"

1.3 Emergency bypass

Only use AI without consent if:

- Patient is unconscious AND
- 3 doctors agree it's life-or-death

2. **Fighting Bias**

2.1 **Monthly bias checkups**

Test if AI works equally well for:

- All skin colors

- Elderly vs. young
- Rare diseases

2.2 **Diverse data diet**

For every 100 patient scans used to train AI:

- Minimum 30 from rural clinics
- Minimum 20 from low-income areas

2.3 **Bias alarms**

If AI makes more errors for any group, it automatically shuts off until fixed

3. **Clear Transparency**

3.1 No black boxes

Doctors get simple reports like:

"Why this diagnosis?

- 80%: Tumor shape matches cancer patterns
- 15%: Your patient's family history
- 5%: Uncertainty"

3.2 **Mistake diaries**

Public log of all AI errors updated weekly:

"Jan 5: Misread a wrist fracture (fixed Jan 8)"

3.3 **Open report cards**

Publish yearly grades:

Heart attack predictions:

- Accuracy: 92%
- Fairness score: A-
- Speed: 3x faster than humans"

4. **Accountability**

4.1 **Patient champions**

Each hospital must have:

- 1 nurse focused on AI safety
- 1 patient advocate on the tech team

4.2 **Harm repair fund**

If AI causes harm:

- Hospital pays all costs
- Engineers help fix the problem
- Apology within 48 hours