Part 1: Theoretical Understanding

1. Short Answer Questions

Q1: What is algorithmic bias and can you give two examples of how it shows up in AI systems?

Algorithmic bias happens when an AI system spits out unfair results because of bad data or wonky design choices. It is like when the system picks up on real-world prejudices or gets trained on lopsided info and then makes decisions that are not fair to everyone. This can mess with people's lives if it's not caught.

- Example 1: A company builds an AI to screen job applicants but trains it on old resumes where most hires were men. The AI ends up downgrading women's applications even if they are just as qualified. This happened with a real hiring tool Amazon tried a while back.
- Example 2: Some facial recognition tech struggles with darker skin tones because the training data mostly had lighter-skinned faces. This leads to more mistakes like misidentifying people which can cause serious problems in things like security checks.

Q2: What is the difference between transparency and explainability in AI? Why do they both matter?

Transparency is about being open with how an AI works. It's like letting people peek under the hood to see what data and algorithms are being used. You're showing the whole process so everyone knows what's going on. Explainability is more about making the AI's decisions easy to understand. Like if the AI says no to a loan you get a clear reason why in plain language not just tech jargon.

Both are super important. Transparency builds trust because people can see the system isn't hiding anything sketchy. Explainability helps regular folks understand what's happening so they can question or fix mistakes. Together they make AI feel fair and less like a mysterious black box.

Q3: How does GDPR affect AI development in the EU?

GDPR is the EU's big data protection law and it shakes up AI development in a few ways. It says you need clear permission from users before using their personal info which limits how much data you can grab for training AI. It also pushes for privacy-by-design so developers have to build systems that protect user data from the start. Plus users have a right to know why an AI made a decision like if it rejected their credit application. This forces developers to make AI more explainable. GDPR makes building ethical AI tougher and pricier but it is all about keeping things fair and safe for users.

2. Ethical Principles Matching

Here’s how the principles line up with their definitions:

- A) Justice: Making sure AI’s benefits and risks are shared fairly across everyone.
- B) Non-maleficence: Ensuring AI doesn’t hurt people or society.
- C) Autonomy: Respecting people’s rights to control their own data and choices.
- D) Sustainability: Building AI that’s eco-friendly and doesn’t wreck the planet.

Part 2: Case Study Analysis

Case 1: Biased Hiring Tool

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

Tasks:

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

The bias in Amazon’s AI hiring tool came from its training data. The system was built using resumes from past hires, which were mostly men since tech jobs historically leaned male. The AI learned to favor patterns in those resumes, like specific keywords or experiences tied to male candidates, and ended up downgrading women’s applications. For example, it might have penalized terms like “women’s leadership group” or undervalued skills from female-dominated fields. The model design didn’t catch this issue because it wasn’t set up to check for gender fairness, so it just amplified the existing bias in the data.

2. Propose three fixes to make the tool fairer.

First, clean up the training data. Use a more diverse set of resumes that includes equal numbers of qualified men and women. Remove gendered terms or identifiers to stop the AI from picking up on biased patterns. Second, tweak the model to include fairness checks. Add algorithms that test for gender bias during training, like ensuring the AI doesn’t favor one group over another. Third, involve humans in the loop. Have HR folks review the AI’s top picks to catch any unfair decisions and adjust the system based on their feedback. This mix of better data, smarter design, and human oversight can make the tool way fairer.

3. Suggest metrics to evaluate fairness post-correction.

To check if the tool is fairer, use the disparate impact ratio, which compares how often the AI selects men versus women. A ratio close to 1 means it’s treating both groups equally. Another good metric is the equal opportunity difference, which checks if qualified candidates from both genders have the same chance of getting picked. Finally, track the false positive rate, like how

often the AI wrongly rejects qualified women compared to men. These metrics will show if the fixes are actually leveling the playing field.

Case 2: Facial Recognition in Policing

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

Tasks:

1. Discuss ethical risks (e.g wrongful arrests, privacy violations).

Facial recognition systems that misidentify minorities can cause some serious harm. Wrongful arrests are a big risk if the AI flags the wrong person as a suspect, innocent people, especially minorities could end up in jail or harassed by police. This erodes trust in law enforcement and ruins lives. Privacy violations are another issue. These systems often pull data from public cameras or databases without clear consent, which feels like a major invasion, especially for communities already over-policed. Plus, the tech can reinforce stereotypes by over identifying minorities as threats which deepens systemic bias. These risks make it critical to handle facial recognition carefully.

2. Recommend policies for responsible deployment.

To make facial recognition safer start with strict rules on data use. Only use diverse high quality datasets for training with equal representation of all racial groups to cut down on misidentifications. Next require transparency police departments should publicly explain how the tech is used what data it pulls and how decisions are made. Third, set up independent audits. Have outside experts regularly check the system for bias and accuracy, sharing results with the public. Finally, limit its use to low-stakes situations, like finding missing persons, not high-risk stuff like arrests without human double-checking. These policies can help balance the tech’s benefits with its risks.

Part 4: Ethical Reflection

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

Answer:

For my future project I am hyped to build a mobile app that uses AI to suggest workout plans based on a user's fitness goals and health data. I want it to be dope for everyone but also super ethical. To make sure it follows AI ethics principles I will focus on a few key moves. First I will use diverse data to train the AI. I will pull health and fitness info from people of all ages body types and backgrounds so the app doesn't just cater to one group like super fit dudes. This way it won't push unrealistic plans on folks who don't fit that mold.

Next I will keep things transparent. The app will tell users how it picks their workouts like if it is based on their heart rate or goals. I will also make it explainable by giving simple reasons for suggestions like "This plan boosts cardio because you said you want better stamina." Third I will respect user control. People can opt out of sharing sensitive data like weight or medical history and still get a solid plan. Finally I will test for fairness using tools like AI Fairness 360 to check if the app treats all users equally. If I find bias like favoring certain groups I will tweak the data or model to fix it. I will also do regular checks to keep it fair over time. By focusing on diversity transparency and user choice my app will be ethical trustworthy and actually help people get fit.

Bonus Task: Policy Proposal for Ethical AI Use in Healthcare

Ethical AI Guidelines for Healthcare

Our team put together these guidelines to make sure AI in healthcare stays ethical and benefits patients without causing harm. We are focusing on consent bias and transparency to keep things legit.

Patient Consent Protocols
Patients come first so any AI system like those predicting diagnoses or treatment plans needs clear consent. Hospitals should use plain language to explain what data the AI uses how it makes decisions and how it affects care. For example if an AI flags a patient for cancer screening they should know their records were analyzed and why. Patients must have the right to say no to data use without losing access to treatment. Consent forms should be easy to read available in multiple languages and offered online or in-person for accessibility.

Bias Mitigation Strategies
Bias in AI can mess up patient care so we need to stop it early. Developers should train AI on diverse datasets that include patients from different races genders ages and income levels. For instance an AI predicting heart attack risk should not work better for one group because of lopsided data. Tools like AI Fairness 360 can audit for bias checking metrics like equal accuracy across groups. If bias shows up developers should reweight data or use fairness algorithms like debiasing models. Hospitals should also set up diverse ethics boards to review AI systems yearly and catch any unfair patterns.

Transparency Requirements

Trust is huge so healthcare AI needs to be an open book. Providers should explain AI decisions in simple terms like why an AI suggested a specific drug. They should also share what data the AI uses and its limits. Hospitals must publish annual reports on AI performance including accuracy and fairness stats. This lets patients and doctors question decisions and keeps the system accountable. These steps ensure AI in healthcare is fair safe and trusted.