

Question 2

Ethics and AI Harm

Allocational Harm

Application Example: Allocational harm occurs when an AI system unfairly distributes resources or opportunities across groups. A hypothetical example is an AI-powered hiring system used by a large corporation to screen job applicants. The system evaluates resumes and assigns scores to predict candidate suitability. If trained on historical hiring data that reflects past biases (e.g., favoring male candidates for technical roles), the model may assign lower scores to qualified female or minority candidates, resulting in their exclusion from interviews. This perpetuates gender or racial disparities in hiring, denying opportunities to deserving candidates.

Harm Mitigation Strategies:

Collecting Better Data:

- **Strategy:** Curate a more diverse and representative training dataset that includes resumes and hiring outcomes from a balanced pool of candidates across gender, race, and other demographics. For example, actively source data from underrepresented groups or use synthetic data to augment the dataset, ensuring equitable representation.
- **Impact:** By training the model on data that reflects fair hiring practices, the AI system learns to evaluate candidates based on qualifications rather than historical biases. This reduces the likelihood of disproportionately rejecting qualified candidates from underrepresented groups, mitigating allocational harm.
- **Example:** Partner with organizations focused on diversity in tech to collect resumes from a wider applicant pool, ensuring the training data includes equal proportions of male and female candidates with similar qualifications.

Modifying the Training Process:

- **Strategy:** Incorporate fairness-aware algorithms during training, such as adversarial training or fairness constraints (e.g., demographic parity or equal opportunity). For instance, use a regularizer in the loss function to penalize differences in selection rates across groups, ensuring the model treats all demographic groups equitably.

- **Impact:** Modifying the training process to prioritize fairness ensures the model optimizes for both accuracy and equity, reducing disparities in hiring outcomes. This directly addresses allocational harm by promoting equal opportunity for all candidates.
- **Example:** Implement an adversarial training setup where a secondary network is trained to detect bias in the hiring model's predictions (e.g., correlation with gender). The primary model is adjusted to minimize this bias while maintaining predictive performance.