Part 4: Ethical Reflection

Reflection on a Personal Project: Adhering to Ethical AI Principles

In one of my recent projects, I developed an AI-powered tool to detect weapons in video footage using YOLOv8, with the aim of improving community safety in urban areas. Although the technical execution was exciting, I recognized early on that without ethical grounding, such technology could unintentionally cause harm—such as surveillance overreach or misidentification.

To ensure the project aligns with ethical AI principles, I've adopted a proactive approach:

1. **Human-Centered Design**: I involved community feedback at the concept stage to understand their privacy expectations and safety concerns. In the future, I will expand this to include diverse groups to avoid designing with unexamined bias.

2. **Transparency**: I documented how detections are made, which model was used, and its limitations. Future iterations will feature explainable AI elements, making it easier for non-technical users to understand the decision process.

3. **Bias Mitigation**: I plan to fine-tune the model with a diverse, context-specific dataset—particularly one reflecting Kenyan urban environments—to reduce false positives and improve fairness.

4. **Accountability**: I implemented logging of detection events and access history. Moving forward, I will include human-in-the-loop verification before any action is taken based on AI recommendations.

5. **Data Privacy**: No personal identification is stored, and footage is processed locally or securely, adhering to data protection standards. I aim to integrate user consent modules for live deployment.

Through regular audits and alignment with frameworks like the UNESCO Recommendation on the Ethics of Artificial Intelligence and Kenya's Data Protection Act (2019), I intend to scale this project responsibly.

Bonus Task: Policy Proposal — 1-Page Guideline for Ethical AI Use in Healthcare

Title: Ethical AI Use in Healthcare — Guiding Principles and Protocols

Objective:
Ensure the ethical development, deployment, and governance of AI systems in healthcare settings to protect patient rights, promote equity, and enhance healthcare delivery.

1. Patient Consent Protocols

- **Informed Consent**: All patients must be informed in clear, local language when their health data is being used to train or feed AI systems.
- **Voluntary Participation**: Patients have the right to opt in or out without penalty. Consent should be revocable at any time.
- **Data Minimization**: Only essential data should be collected. De-identified data must be used whenever possible to protect privacy.

2. Bias Mitigation Strategies

- **Inclusive Data Collection**: Training datasets must include diverse demographic groups (e.g., age, gender, ethnicity, geography) to prevent underrepresentation.
- **Bias Audits**: Conduct regular model audits using fairness metrics (e.g., equalized odds, demographic parity) to identify and address disparities.
- **Continual Learning**: Continuously monitor AI performance in real-world settings and retrain models to correct emerging biases.

3. Transparency Requirements

- **Model Explainability**: AI tools must provide understandable explanations of outputs, especially when influencing diagnosis or treatment.
- **Documentation**: Developers must publish clear documentation detailing model objectives, training data sources, performance metrics, and limitations.
- **User Training**: Healthcare workers must receive training on how the AI system works, its benefits, risks, and how to override or question decisions.
- **Public Disclosure**: Report all AI use cases to a central registry accessible to patients and the public for accountability.