Ethical Reflection

Personal project reflection
For a hypothetical student-recommendation system I build: I will ensure ethical AI by (1) collecting consented, minimal data and documenting provenance; (2) performing fairness audits (disaggregated metrics) during development; (3) implementing transparency (what data features are used; provide explanations for recommendations and an appeal pathway); (4) including human oversight for sensitive recommendations and monitoring model drift; and (5) minimizing environmental impact through efficient model choices and scheduled retraining only when necessary.

Guideline for ethical AI in healthcare

Ethical AI Use in Healthcare 1-Page Guideline

Purpose & Scope
Covers deployment and use of AI tools that influence patient care decisions, diagnostics, triage, and resource allocation.

1. Patient Consent & Data Governance

- Obtain informed consent for using patient data in model training and inference when feasible. Clearly state purposes, risks, and retention periods.
- Use data minimisation: collect only necessary fields.
- Use robust de-identification and encryption for stored data; restrict access by role.
- Maintain audit logs of data access and model inference for accountability.

2. Bias Mitigation Strategies

- **Data review:** Assess representativeness across age, sex, race/ethnicity, socioeconomic status. Log known gaps.
- **Preprocessing:** Use reweighing or balancing when necessary; remove identified proxies for protected attributes.
- **Modeling:** Prefer interpretable models for high-stakes decisions where possible. If using complex models, apply fairness-aware training (adversarial debiasing, constrained optimization).
- **Evaluation:** Report disaggregated performance metrics (sensitivity, specificity, FPR, FNR) and calibration across groups prior to deployment.
- **Monitoring:** Continuous post-deployment monitoring for performance drift and fairness degradation.

3. Transparency & Explainability

- Publish model intent, use-cases, major features, and limitations in clinician-facing documentation.
- Provide patient-appropriate explanations for automated recommendations; clinicians must be able to query feature contributions.
- Maintain versioning and changelogs; any retraining requires re-evaluation.

4. Clinical Oversight & Human-in-the-loop

- AI outputs should support, not replace, clinical judgment. Require clinician review and sign-off for critical decisions.
- Define escalation paths for conflicts between model and clinician judgment.

5. Privacy, Safety & Regulatory Compliance

- Ensure compliance with local healthcare privacy laws (e.g., GDPR, HIPAA).
- Conduct Data Protection Impact Assessments (DPIAs) and clinical safety risk assessments.
- For high-risk systems, pursue independent third-party audits and validation studies.

6. Patient Redress & Accountability

- Provide a clear mechanism for patients to question or appeal AI-influenced decisions.
- Assign organizational ownership for model governance and address harm mitigation fast.

7. Environmental Sustainability

- Prefer efficient models; schedule retraining and experimentation mindfully to reduce energy use.

8. Deployment Controls

- Limit usage to approved clinical contexts, ensure logging, rollback capability, and clear stop-criteria when performance fails.