MODULE 6 : ETHICS AND FUTURE OF AI

Module 6: Ethics and Future of AI — 5 Lessons

Designed to move beyond slogans into concrete practices you can apply when designing, deploying, and governing AI systems.

Lesson 1: Fairness from Data to Decisions

Measuring and mitigating bias in real workflows

Bias creeps in through datasets (sampling, labeling), features (proxies for protected attributes), and decision thresholds (different error costs across groups). A practical workflow is: define harm and stakeholders → choose fairness metrics aligned to context → run disaggregated evaluation (by subgroup) → mitigate via data balancing/reweighting, algorithmic changes (e.g., constraints for equalized odds), or post-processing (group-aware thresholds) → monitor drift over time. Remember that metrics trade off: demographic parity equalizes positive rates, while equalized odds equalizes error rates—pick based on the real-world duty of care (e.g., screening vs. adjudication). Document decisions and residual risks.

Key terms: demographic parity, equalized odds, calibration, subgroup analysis, post-processing

Lesson 2: Privacy & Data Governance by Design

Protecting people while preserving utility

Build around data minimization (collect only what you need), purpose limitation (no surprise uses), and secure lifecycle (ingestion → storage → access → deletion). Technical safeguards include differential privacy (adding calibrated noise so single records have limited influence), federated learning (train where data lives), and secure aggregation or confidential computing to reduce exposure. Beware "anonymous" data—linkage attacks can re-identify rows. Track lineage and consent; log who accessed what and why. When generating synthetic data, validate that it doesn't memorize individuals (membership inference tests). Make privacy impact assessments routine, not exceptional.

Key terms: differential privacy (ε), federated learning, secure aggregation, data minimization, lineage

Lesson 3: Transparency, Explainability & Accountability

Making systems legible and auditable

Transparency is more than a one-time disclosure. Publish Model Cards (intended use, training data sources, metrics by subgroup, known limitations) and Datasheets for Datasets (collection process,

consent, licensing, caveats). Maintain decision logs (inputs, versioned model, thresholds) for auditability and incident response. Use explainability responsibly: local methods (e.g., SHAP/LIME) can aid debugging, but don't over-promise causal truth; pair explanations with user education and uncertainty ranges. Define escalation paths: who pauses a model, how rollbacks work, and what remediation looks like for affected users.

Key terms: model card, datasheet, audit trail, SHAP/LIME, uncertainty disclosure

Lesson 4: Safety, Alignment & Misuse Prevention

From red-teaming to runtime safeguards

Operational safety blends pre-deployment testing with live defenses. Red-team models to probe for harmful content, data exfiltration, jailbreaks, prompt injection, or privacy leaks. Align models with policy via instruction tuning and refusal/deflection strategies; layer content filters, rate limiting, and abuse detection in production. For connected systems and agents, constrain tools (principle of least privilege), validate outputs (grounding/verification), and add human-in-the-loop for high-impact actions. Continuously monitor for drift and incidents; create a lightweight Safety Review checklist for every change (new data, new prompt, new integration).

Key terms: red-teaming, alignment, content filtering, prompt injection, least privilege

Lesson 5: The Near Future of AI—Trends You Can Use

What to build, how to evaluate, and where risks shift

Three impactful shifts: (1) Retrieval-Augmented Generation (RAG) to ground outputs in your sources, reducing hallucinations and enabling citations; invest in document chunking, metadata, and evaluation beyond string-match. (2) On-device/edge models for privacy, lower latency, and resilience; plan for model quantization and fallback paths. (3) Agentic workflows that chain tools and tasks; require guardrails, state inspection, and sandboxed execution. Expect stronger sustainability pressures—optimize with distillation, quantization, caching, and scheduled batch jobs. Evaluation must evolve: add task-specific checklists, human ratings, and safety benchmarks alongside accuracy.

Key terms: RAG, grounding, quantization, distillation, agent sandboxing

Quick Check for Lesson 1: Fairness from Data to Decisions

Which statement best distinguishes demographic parity from equalized odds?
• A. Demographic parity equalizes error rates; equalized odds equalizes selection rates
• B. Demographic parity equalizes selection rates; equalized odds equalizes error rates ☑
• C. Both require identical ROC curves across groups
• D. Both guarantee optimal accuracy

Quick Check for Lesson 2: Privacy & Data Governance by Design

What does differential privacy aim to guarantee?
• A. Encrypted storage at rest
• B. Individual records have limited influence on outputs, even if attacked ☑
• C. Models never overfit
• D. Data cannot be shared across organizations

Quick Check for Lesson 3: Transparency, Explainability & Accountability

Which item is most appropriate to include in a Model Card?
• A. The SHA of your Git repository only
• B. Intended use, subgroup metrics, and known limitations ☑
• C. Raw training data dumps
• D. Proprietary feature weights

Quick Check for Lesson 4: Safety, Alignment & Misuse Prevention

What is the primary purpose of red-teaming an AI system?
• A. Speeding up inference
• B. Stress-testing for unsafe behaviors and vulnerabilities before deployment ☑
• C. Reducing compute cost
• D. Increasing dataset size

Quick Check for Lesson 5: The Near Future of AI—Trends You Can Use

What is the main goal of Retrieval-Augmented Generation (RAG)?
• A. Speed up training by pruning layers
• B. Ground model outputs in external, trusted sources at query time ☑
• C. Replace the need for vector databases
• D. Guarantee identical outputs every run

Module Summary

Ethical and future-ready AI hinges on disciplined practice. Start by defining harm and stakeholders, then measure fairness with subgroup-aware metrics aligned to your setting, acknowledging trade-offs between parity and error rates. Protect people through privacy-by-design—collect less, constrain purposes, and use differential privacy, federated learning, and secure aggregation where appropriate. Make systems legible with model cards, datasheets, audit trails, and honest uncertainty disclosures; pair explanations with clear escalation paths. Treat safety as continuous: red-team regularly, align models with policy, restrict privileges, and monitor incidents. Looking forward, ground generation with RAG, move sensitive inference to the edge when feasible, and sandbox agent actions. Evaluate beyond single metrics—combine task success, human feedback, and safety checks—so your systems remain robust as data, users, and contexts change.