

# Module V: AI Ethics (Case Studies)

 by Mrs. Purva D Thakare

# Open Research Problems in AI Ethics

## Near-Term Ethical Problems

These are problems happening **right now or very soon** because of AI.

Examples include:

- **Bias:** AI treats some groups unfairly.
- **Privacy issues:** People's personal data gets misused or leaked.
- **Wrong decisions:** AI makes mistakes in important areas (like healthcare or hiring).
- **Misinformation:** AI spreads fake news or false information.
- **Misuse:** People use AI for harmful purposes.
- **No accountability:** It's not clear who is responsible when AI causes harm.

## Long-Term Ethical Problems

These are problems that might **develop slowly over time** and affect society in a big way.

Examples include:

- **Power concentration:** A few companies or countries control most of the AI technology.
- **Misaligned superintelligent AI:** Very advanced AI may not follow human values.
- **Erosion of social norms:** Over time, AI could change how people think, work, and interact in harmful ways.

### Near-Term (Current and Practical Issues)

#### • Measuring and Reducing Bias

Problem: AI systems can be unfair to some groups. It's hard to find and fix bias in different situations.

Research focus:

How to test for bias in any domain

How to audit AI systems fairly

Finding what causes unequal results

#### • Making AI Explanations Easy to Understand

Problem: AI decisions are often confusing for regular people.

Research focus:

How to make explanations simple and useful

How to create explanations for different types of users

#### • Accountability and Transparency

Problem: Once an AI system is deployed, it's hard to know how it behaves or who is responsible for its actions.

Research focus:

Setting clear rules for AI audits

Building systems that record how AI makes decisions

Tracking where AI models come from and how they change

#### • Balancing Privacy and Usefulness

Problem: Protecting user data while still making AI systems accurate and helpful.

Research focus:

Creating privacy-protecting methods that still work well

Studying privacy in systems that share data (like federated learning)

#### • Fighting Misinformation and Fake Content

Problem: AI can create fake news, deepfakes, or misleading information.

Research focus:

Adding invisible "marks" to AI-generated content to trace its source

Detecting fake or AI-generated media

Building tools to identify when AI is used to spread false information

#### • Sharing Decisions Between Humans and AI

Problem: Deciding when to trust AI and when humans should stay in control. Research focus:

Setting rules for when AI should make decisions

Designing user interfaces for better teamwork between humans and AI

Helping humans build the right amount of trust in AI systems

### Long-Term (Future and Big-Picture Issues)

#### • Aligning AI with Human Values

Problem: Making sure advanced AI systems respect human values and cultures.

Research focus:

Getting feedback from people at scale

Defining what "aligned" behaviour means

Combining different human preferences fairly

#### • Avoiding Power Concentration

Problem: A few companies or countries might control most powerful AI systems.

Research focus:

Creating fair rules and governance for AI

Designing systems for shared and transparent control

Ensuring competition and public benefit

#### • Preparing for Job and Social Changes

Problem: Automation may replace jobs and change how society works. Research focus:

Designing policies for worker retraining

Building strong social safety systems

Supporting people in adapting to new roles

#### • Long-Term Safety and Global Risks

Problem: Preventing very rare but extremely dangerous AI outcomes in the future. Research focus:

Predicting and studying possible long-term risks

Creating global rules for AI safety

Working together internationally on AI governance

# Challenges / opportunities / possible approaches

- **Challenges:** measurement difficulties, adversarial adaptation, changing social norms, cross-jurisdiction regulation, incentive misalignment (profit vs public good).
- **Opportunities:** improve fairness and access, augment human decision making, scale beneficial services (healthcare, education), transparent public services.
- **Approaches:** interdisciplinary teams (tech + social science + law), participatory design, impact assessments, continuous monitoring, regulatory sandboxes, incentives for open models and reproducible audits.

# Societal issues for AI in medicine

- **Major concerns:** biased training data → unequal treatment; opaque clinical recommendations; data privacy and consent; liability for errors; workflow disruption for clinicians.
- **Approaches:** clinical trials for algorithms, clear human-in-the-loop rules, provenance/traceability for training data, patient-centred consent models, post-deployment monitoring, explainability tailored for clinicians and patients.
- **Research needs:** benchmarks for clinical fairness, methods to combine causal medical knowledge with data-driven models, standards for regulatory approval of AI tools.

# Decision-role in industries (how AI should be used)

- **Decision taxonomy:** advisory (recommendation), assisted (high autonomy but human final sign-off), automated (machine makes decision), supervisory (human monitors many automated actions).
- **Design rules:** critical/high-risk decisions → keep human final authority; low-risk, high-scale operations → allow automation with monitoring; ensure fallbacks and escalation paths.
- **Industrial best practices:** risk categorization, rigorous testing in production environments, continuous evaluation metrics (fairness, accuracy, reliability), clear audit trails and governance committees.

# Relation between AI and democracy

- **Risks:** manipulation (microtargeting, deepfakes), asymmetric information, concentrated influence by private platforms, automated surveillance weakening civil liberties.
- **Safeguards:** transparency requirements for political ads, provenance/watermarking for synthetic media, platform accountability, public digital literacy programs, independent audits of civic systems.
- **Research questions:** detection of political manipulation at scale, algorithmic transparency that balances privacy and public interest, governance designs to preserve pluralism.

# National & international strategies on AI (what good strategies must include)

- **Core elements:** safety and standards, R&D investment (public interest tech), workforce transition programs, data governance and privacy regimes, international coordination on high-risk AI, support for civil society oversight.
- **Operational tools:** regulatory sandboxes, model registries, mandatory impact assessments for high-risk systems, export controls for dangerous capabilities, public funding for verification/audit bodies.
- **Research/policy gap:** interoperable international standards and enforcement mechanisms.

# Benefits of ethical AI

- Improved trust & adoption, reduced harm and legal risk, fairer access to services, better long-term social outcomes, smoother regulatory compliance, and increased resilience to misuse.