

ETHICAL AND RESPONSIBLE AI

Code: CS33223CS	Course Type: Elective	Semester: II
Evolution Scheme (TA-MSE-ESE): 20-30-100	Total Marks: 150	L-T-P (Credits): 3-0-0 (3)

Pre-requisites

- Basics of Artificial Intelligence and Machine Learning
- Fundamental knowledge of Data Science
- Basic understanding of Probability and Statistics
- Awareness of societal and ethical issues in technology
- Programming knowledge (Python preferred)

Course Objectives

1. To understand ethical challenges and societal impacts of Artificial Intelligence systems.
2. To analyze issues related to bias, fairness, accountability, and transparency in AI.
3. To study data privacy, governance, and regulatory frameworks (e.g., GDPR, AI Act).
4. To design trustworthy, explainable, and responsible AI systems.
5. To evaluate risk, compliance, and ethical decision-making in real-world AI applications.

Course Outcomes (COs)

After completing the course, students will be able to:

CO1: Understand ethical principles, risks, and societal implications of AI systems.

CO2: Analyze bias, fairness, and risk in machine learning models and propose mitigation techniques.

CO3: Apply data privacy, governance, and regulatory frameworks in AI system design.

CO4: Design explainable, transparent, and trustworthy AI solutions.

CO5: Evaluate and implement responsible AI practices in real-world applications.

Course Articulation Matrix

COs	PO1 (Knowledge)	PO2 (Analysis)	PO3 (Design)	PO4 (Investigation)	PO5(Modern Tools)	PO6 (Research/ Innovations)
CO1	2	2	1	1	-	3
CO2	3	3	2	2	2	3
CO3	3	3	2	3	2	3
CO4	3	3	3	3	3	3
CO5	3	3	3	3	3	3

3 → Substantial; 2 → Moderate ; 1 → Slight

Syllabus

UNIT I: Foundations of Ethical AI

Introduction to Ethical and Responsible AI, Motivation and societal impact of AI, AI ethics principles: fairness, accountability, transparency, responsibility, AI-driven decision-making and organizational impact, Ethical theories (utilitarianism, deontology, virtue ethics), AI risks and unintended consequences

UNIT II: Bias, Fairness, and Risk in AI

Bias of AI: Types of bias in AI (data bias, algorithmic bias, societal bias), Sources and impact of bias in ML models, Fairness metrics and evaluation techniques, Bias mitigation strategies, AI risks and failure cases, Case studies of biased AI systems

UNIT III: Privacy, Security, and Governance

Data privacy concepts and challenges, Data governance and ethical data usage, Privacy-preserving techniques; Regulatory frameworks: GDPR, EU AI Act, Global AI policies; Risk management frameworks (e.g., NIST AI RMF), Security threats and adversarial risks in AI

UNIT IV: Explainability, Transparency, and Trustworthy AI

Explainable AI (XAI): importance and techniques, Model interpretability vs explainability, Transparency in AI systems, Human-in-the-loop systems, Trustworthy AI frameworks, Responsible AI lifecycle (design, deployment, monitoring)

UNIT V: Responsible AI in Practice

Ethical AI in applications: Healthcare, Finance, Autonomous systems; AI governance and accountability mechanisms, Ethical auditing and compliance, Responsible AI development lifecycle; Emerging trends: AI alignment, sustainability, AI for social good, Case studies and real-world implementation challenges

Textbook

1. Virginia Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, Springer, 2019.
2. Ursula Ferretti et al., *Responsible Artificial Intelligence: Ethics and Regulation*, Springer, 2023.

Reference Book

1. Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of AI*, Yale University Press, 2021.
2. Brian Christian, *The Alignment Problem: Machine Learning and Human Values*, W.W. Norton, 2020.