

# PRASAD V. POTLURI SIDDHARTHA INSTITUTE OF TECHNOLOGY

(Autonomous) Kanuru, Vijayawada-520007

## Department of Computer Science Engineering (Data Science)

IV B. Tech – I Semester

### Responsible AI

|                                       |           |                                |             |                      |           |
|---------------------------------------|-----------|--------------------------------|-------------|----------------------|-----------|
| <b>Course Code</b>                    | 23AM4701A | <b>Year</b>                    | IV          | <b>Semester</b>      | I         |
| <b>Course Category</b>                | PE-IV     | <b>Branch</b>                  | CSE (AI&ML) | <b>Course Type</b>   | Theory    |
| <b>Credits</b>                        | 3         | <b>L-T-P</b>                   | 3-0-0       | <b>Prerequisites</b> | AI and ML |
| <b>Continuous Internal Evaluation</b> | 30        | <b>Semester End Evaluation</b> | 70          | <b>Total Marks</b>   | 100       |

| Course Outcomes                                                   |                                                                                                                                                        |           |
|-------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| Upon Successful completion of course, the student will be able to |                                                                                                                                                        |           |
| <b>CO1</b>                                                        | Describe the concepts, principles, ethical concerns, fairness, explainability, privacy and security of AI in the development of responsible AI systems | <b>L2</b> |
| <b>CO2</b>                                                        | Use fairness metrics, bias assessment methods, and mitigation techniques in AI systems to identify and reduce bias in AI-based decisions               | <b>L3</b> |
| <b>CO3</b>                                                        | Apply explainability and interpretability techniques to interpret and justify predictions of AI models.                                                | <b>L3</b> |
| <b>CO4</b>                                                        | Analyze security, privacy, and ethical challenges in AI systems to assess deployment risks and support responsible adoption.                           | <b>L4</b> |

| Contribution of Course Outcomes towards achievement of Program Outcome & Strength of correlation (3: High, 2: Medium, 1: Low) |     |     |     |     |     |     |     |     |     |      |      |      |      |
|-------------------------------------------------------------------------------------------------------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|------|------|------|
|                                                                                                                               | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PSO1 | PSO2 |
| <b>CO1</b>                                                                                                                    | 3   |     |     |     |     |     | 3   |     |     |      |      |      |      |
| <b>CO2</b>                                                                                                                    | 3   |     |     |     |     | 3   | 3   |     |     |      |      | 1    | 1    |
| <b>CO3</b>                                                                                                                    | 3   |     |     |     |     |     |     |     |     |      |      |      |      |
| <b>CO4</b>                                                                                                                    |     | 3   |     | 3   |     | 3   | 3   |     |     |      |      | 1    | 1    |

| Syllabus |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |                    |
|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------|
| Unit No  | Contents                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | Map ped CO         |
| I        | <b>INTRODUCTION</b> : Brief Overview of AI and its Potential: Foundations of AI: From Concept to Reality; AI in Action: A Multifaceted Landscape; The Promise of AI: Unlocking Boundless Potential; Navigating the AI Frontier, Importance of Responsible AI: Ethics in the age of AI, Mitigating Bias and Discrimination, Privacy in the age of Surveillance, Human Centric design, Ethics in AI Governance, Core Ethical principles : Bias and Fairness, Transparency and Explainability, Privacy, and Security, Robustness and Reliability.[Textbook 1]                                                                                                                                                                | CO1                |
| II       | <b>BIAS AND FAIRNESS:</b> Understanding Bias in Data and Models, Importance of Understanding Bias, How Bias can impact decision making Processes, Types of bias: Data Bias; Model Bias; Social Bias, Examples of Real-world Cases: Amazon's Gender-Biased Hiring Tool; Racial Bias in Criminal Risk Assessment; Google Photos' Mislabeling Incident ; Biased Loan Approval Models ; Facial Recognition and Racial Bias ,Techniques to Detect and Mitigate Bias: Techniques to detect Bias, Techniques to mitigate Bias. [Textbook 1]                                                                                                                                                                                      | CO1, CO2           |
| III      | <b>TRANSPARENCY AND EXPLAINABILITY:</b> Transparency, Explainability, Importance of Transparency and Explainability: trust; Accountability; Ethical Considerations; Informed Decisions; Error Detection and Management, Real World Examples of the Impact of Transparent AI, Methods for Achieving Explainable AI: Explanation Methods for Interpretable Models, Generating Feature Importance Scores and Local Explanations: Generating Feature Importance Scores, Generating Local Explanations, Creating Saliency Maps, Overview of Tools and Libraries for AI Model Transparency: InterpretML, SHAP, LIME, TensorBoard, XAI(Explainable AI) Libraries, ELSI5. [Textbook 1]                                            | CO1, CO3           |
| IV       | <b>PRIVACY AND SECURITY:</b> Privacy Concerns in AI, Potential Threats to Privacy: Data Breaches and Unauthorized Access; Misuse of Personal Data by AI Models; Inadvertent Sharing of Sensitive Information, Privacy Attacks in AI Models: Data Re-identification; Inference Attacks; Membership Inference Attacks; Model Inversion Attacks, Security Concerns in AI: Potential Threats to Security: Adversarial Attack; Data Poisoning; Model Inversion and Extraction; Evasion Attack; Backdoor Attack , Can AI systems be Ethical?[TextBook 2]: What Is an Ethical Action?, Approaches to Ethical Reasoning by AI, Designing Artificial Moral Agents, Implementing Ethical Deliberation, Levels of Ethical Behaviour. | CO1, CO4           |
| V        | <b>CASE STUDIES:</b> Unfair and Illegal Discrimination: Cases of AI-Enabled Discrimination: Gender Bias in recruiting Tools; Discrimination on the Basis of Skin Colour , Privacy: Cases of Privacy Violations Through AI: Genetic Privacy; Biometric Surveillance, Right to Life, Liberty and Security of Persons : Cases of AI Adversely Affecting the Right to Life, Liberty and Security : Fatal Crash Involving a Self-driving Car, Smart Home Hubs Security Vulnerabilities [TextBook 3]                                                                                                                                                                                                                            | CO1, CO2, CO3, CO4 |

| <b>Learning Resources</b>                     |                                                                                                                                                                                     |
|-----------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Text Books</b>                             |                                                                                                                                                                                     |
| 1.                                            | Introduction to Responsible AI: Implement Ethical AI Using Python, Avinash Manure, Shaleen Bengali, Saravanan S, 1st Edition, 2023, Apress.                                         |
| 2.                                            | Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way, Virginia Dignum, 1st edition, 2019, Springer Cham                                              |
| 3.                                            | Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges, Bernd Carsten Stahl, Doris Schroeder, Rowena Rodrigues, 1st Edition, 2023, Springer. |
| <b>References</b>                             |                                                                                                                                                                                     |
| 1.                                            | Human Compatible: Artificial Intelligence and the Problem of Control, Stuart J. Russell, 1st Edition, 2019, Viking / Penguin Press                                                  |
| 2.                                            | Trustworthy AI, Beena Ammanath, 1st edition, 2022, O' Reilly.                                                                                                                       |
| 3.                                            | Interpretable Machine Learning, Christoph Molnar ,1st edition, 2019, Lean Publishing                                                                                                |
| 4.                                            | <i>Fairness and Machine Learning: Limitations and Opportunities</i> , Solon Barocas, Moritz Hardt, and Arvind Narayanan, 1st edition, 2023, MIT Press.                              |
| <b>E-Recourses and other Digital Material</b> |                                                                                                                                                                                     |
| 1.                                            | <a href="https://www.coursera.org/learn/responsible-and-ethical-ai">https://www.coursera.org/learn/responsible-and-ethical-ai</a>                                                   |
| 2.                                            | <a href="https://onlinecourses.nptel.ac.in/noc24_cs132/preview">https://onlinecourses.nptel.ac.in/noc24_cs132/preview</a>                                                           |