SEMESTER S7

RESPONSIBLE ARTIFICIAL INTELLIGENCE

Course Code	**PECST752**	**CIE Marks**	40
Teaching Hours/Week (L: T:P: R)	3:0:0:0	**ESE Marks**	60
Credits	3	**Exam Hours**	2 Hrs. 30 Min.
Prerequisites (if any)	None	**Course Type**	Theory

Course Objectives:

1. To impart the ideas of fairness, accountability, bias, and privacy as fundamental aspects of responsible AI.
2. To teach the principles of interpretability techniques including simplification, visualization, intrinsic interpretable methods, and post hoc interpretability for AI models.
3. To give the learner understanding of the ethical principles guiding AI development, along with privacy concerns and security challenges associated with AI deployment.

SYLLABUS

Module No.	Syllabus Description	Contact Hours
1	**Foundations of Responsible AI :-** Introduction to Responsible AI- Overview of AI and its societal impact; Fairness and Bias - Sources of Biases, Exploratory data analysis, limitation of a dataset, Preprocessing, inprocessing and postprocessing to remove bias.	7
2	**Interpretability and explainability:-** Interpretability - Interpretability through simplification and visualization, Intrinsic interpretable methods, Post Hoc interpretability, Explainability through causality, Model agnostic Interpretation. Interpretability Tools - SHAP (SHapley Additive exPlanation), LIME(Local Interpretable Model-agnostic Explanations)	10
3	**Ethics, Privacy and Security :-** Ethics and Accountability -Auditing AI models, fairness assessment, Principles for ethical practices. Privacy preservation - Attack models, Privacy-preserving Learning, Differential privacy- Working, The Laplace Mechanism, Introduction to	10

	Federated learning. Security - Security in AI Systems, Strategies for securing AI systems and protecting against adversarial attacks	
4	**Future of Responsible AI and Case Studies : -** Future of Responsible AI - Emerging trends and technologies in AI ethics and responsibility. Case Studies - Recommendation systems, Medical diagnosis, Computer Vision, Natural Language Processing.	**9**

Course Assessment Method
(CIE: 40 marks, ESE: 60 marks)

Continuous Internal Evaluation Marks (CIE):

Attendance	Assignment/ Microproject	Internal Examination-1 (Written)	Internal Examination- 2 (Written)	Total
5	**15**	**10**	**10**	**40**

End Semester Examination Marks (ESE)

In Part A, all questions need to be answered and in Part B, each student can choose any one full question out of two questions

Part A	Part B	Total
• 2 Questions from each module. • Total of 8 Questions, each carrying 3 marks **(8x3 =24 marks)**	• Each question carries 9 marks. • Two questions will be given from each module, out of which 1 question should be answered. • Each question can have a maximum of 3 subdivisions. **(4x9 = 36 marks)**	**60**

Course Outcomes (COs)

At the end of the course students should be able to:

Course Outcome		Bloom's Knowledge Level (KL)
CO1	Identify and describe key aspects of responsible AI such as fairness, accountability, bias, and privacy.	**K2**
CO2	Describe AI models for fairness and ethical integrity.	**K2**
CO3	Understand interpretability techniques such as simplification, visualization, intrinsic interpretable methods, and post hoc interpretability.	**K2**
CO4	Comprehend the ethical principles, privacy concerns, and security challenges involved in AI development and deployment.	**K3**
CO5	Understand responsible AI solutions for practical applications, balancing ethical considerations with model performance.	**K3**

Note*: K1- Remember, K2- Understand, K3- Apply, K4- Analyse, K5- Evaluate, K6- Create*

CO-PO Mapping Table (Mapping of Course Outcomes to Program Outcomes)

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	3	3	3									3
CO2	3	3	3									3
CO3	3	3	3									3
CO4	3	3	3									3
CO5	3	3	3									3

Note: *1: Slight (Low), 2: Moderate (Medium), 3: Substantial (High), -: No Correlation*

Text Books				
Sl. No	**Title of the Book**	**Name of the Author/s**	**Name of the Publisher**	**Edition and Year**
1	Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way	Virginia Dignum	Springer Nature	1/e, 2019
2	Interpretable Machine Learning	Christoph Molnar	Lulu	1/e, 2020

Reference Books				
Sl. No	**Title of the Book**	**Name of the Author/s**	**Name of the Publisher**	**Edition and Year**
1	ResponsibleAI Implementing Ethical and Unbiased Algorithms	Sray Agarwal, Shashin Mishra	Springer Nature	1/e, 2021

Video Links (NPTEL, SWAYAM…)	
Module No.	**Link ID**
1	https://youtu.be/3-xhMXeYIcg?si=x8PXrnk0TabaWxQV
2	https://youtu.be/sURHNhBMnFo?si=Uj0iellJs3oLOmDL [SHAP and LIME] https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/ https://shap.readthedocs.io/en/latest/ https://www.kaggle.com/code/bextuychiev/model-explainability-with-shap-only-guide-u-need
3	https://www.youtube.com/live/DA7ldX6OIG4?si=Dk4nW1R1zi_UMG_4
4	https://youtu.be/XlYhKwRLerc?si=IeU7C0BLhwn9Pvmi Case Studies https://www.kaggle.com/code/teesoong/explainable-ai-on-a-nlp-lstm-model-with-lime https://www.kaggle.com/code/victorcampelo/using-lime-to-explaining-the-preditions-from-ml