Atal Bihari Vajpayee Indian Institute of Information Technology & Management, Gwalior

IT304: Trustworthy Artificial Intelligence

Major Examination (Session 2023–24)

Maximum Time: 3 Hours **Max Marks: 45**

Note: All questions are compulsory. Justify your answers with suitable case studies.

1. (a) Define the principles of trustworthy AI. (b) Discuss the role of ethics in AI system design. (7 Marks)
2. (a) What is model robustness? Explain techniques to make AI robust against adversarial attacks. (b) Give one practical example of adversarial input. (8 Marks)
3. (a) Differentiate between explainability and interpretability in AI. (b) Describe one XAI (Explainable AI) technique with example. (8 Marks)
4. (a) What is bias in training datasets? How can it be mitigated? (b) Discuss fairness metrics used in AI evaluation. (7 Marks)
5. (a) Explain the concept of accountability in AI systems. (b) Discuss the importance of human oversight in automated decision making. (8 Marks)
6. Write short notes on any two: (i) Privacy-preserving machine learning (ii) Transparency in deep learning (iii) Societal impacts of untrustworthy AI (7 Marks)