ABHYASMITRA.IN
COMPREHENSIVE UNIVERSITY STUDY NOTES

ARTIFICIAL INTELLIGENCE

Unit V — Planning and Emerging AI Topics

Topics Covered in this Unit

Automated & Classical Planning: Definition, PDDL (Planning Domain Definition Language) & STRIPS Representation

Planning State-Space Search: Forward State Progression, Backward Goal Regression & Planning Graphs (Graphplan)

Hierarchical Task Network (HTN) Planning: Non-Primitive Tasks, Method Decomposition & Partial-Order Planning

Planning Under Uncertainty: Contingent Planning, Conformant Planning & Markov Decision Processes (MDPs)

Philosophical Limits, Ethics & Alignment of AI: Chinese Room, Autonomous Systems & The Alignment Problem

Generative AI & Foundation Models: Transformer Architecture (Self-Attention), Large Language Models (GPT) & Diffusion Models (DALL-E)

Explainable AI (XAI): Interpretability, Post-Hoc Explanations, LIME, SHAP & Gradient Saliency Maps

Federated Learning (FL): Privacy-Preserving Decentralized Machine Learning & The FedAvg Algorithm

Edge AI: On-Device Intelligence, Model Quantization, Pruning, Knowledge Distillation & Ultra-Low Latency

High-Yield University Exam Question Bank with Detailed Model Answers & Unit Summary

Compiled in a structured, easy-to-understand, textbook style format
Specially curated for B.Tech / B.E. / BCA / MCA / B.Sc Computer Science, AI & Data Science Students

COURSE SYLLABUS & MAPPING

[UNIT V SYLLABUS — PRESCRIBED UNIVERSITY CURRICULUM]

This section contains the official syllabus for Unit V. Verify with your university curriculum mapping.

Unit V Syllabus Breakdown:

Unit V Planning and Emerging AI Topics: Automated and Classical Planning, Hierarchical Planning, Planning under Uncertainty, Analysis of Planning Approaches, Limits of AI, Ethics of AI, Future of AI, AI Components, Introduction of Generative AI (GPT, DALL·E), Explainable AI (XAI), Federated Learning, Edge AI

• Course Code: ____________________ • Semester / Year: ____________________

TABLE OF CONTENTS

5.1 Automated & Classical Planning (STRIPS & PDDL)

Definition: Automated Planning
Automated Planning is the branch of AI concerned with the realization of action sequences, typically executed by intelligent agents, autonomous robots, and unmanned vehicles. Classical planning operates under assumptions of determinism, full observability, discrete time, and static environments.

PDDL (Planning Domain Definition Language) Representation

- • **State Representation:** A state is represented as a conjunction of ground, function-free positive literals (e.g., At(Plane1, JFK) ∧ At(Cargo1, SFO) ∧ In(Cargo2, Plane1)). Closed-World Assumption applies (any literal not mentioned is False).
- • **Actions Schema:** Defined by Action Name, Parameter List, Preconditions, and Effects:
  - • **Action:** Fly(p, from, to)
  - • **Preconditions:** At(p, from) ∧ Plane(p) ∧ Airport(from) ∧ Airport(to)
  - • **Effects:** At(p, to) ∧ ¬ At(p, from) (Add-List: At(p, to); Delete-List: At(p, from))

Planning Search Strategies

Search Direction	Operational Mechanism	Advantages & Limitations
Forward State-Space Search (Progression)	Starts at initial state s_0, finds applicable actions whose preconditions are satisfied, applies effects, and searches forward towards goal.	Sound and complete; maintains exact ground state, but suffers from massive branching factor.
Backward Relevant-States Search (Regression)	Starts at goal state g, finds relevant actions whose effects achieve a goal literal, regresses preconditions backward to find predecessors.	Only explores actions relevant to goal (much smaller branching factor), but manipulating set of states is computationally complex.
Planning Graph (Graphplan)	Constructs an alternating graph of Proposition Levels and Action Levels with Mutual Exclusion (Mutex) relations.	Extracts highly accurate domain-independent heuristics (h_max, h_sum) in polynomial time.

5.2 Hierarchical Planning & Planning Under Uncertainty

Definition: Hierarchical Task Network (HTN) Planning
HTN planning decomposes high-level abstract non-primitive tasks (e.g., 'Build House', 'Fly from NYC to London') into successively refined sub-tasks until executable primitive actions are reached, mirroring human top-down organizational problem solving.

Planning Under Uncertainty

- • **Sensorless / Conformant Planning:** Used when the agent has zero sensors (fully unobservable). Finds a single open-loop sequence of actions that achieves the goal in all possible initial belief states.
- • **Contingent / Conditional Planning:** Used in partially observable environments with sensing actions. Generates planning trees containing conditional branches (IF sense_result THEN plan_A ELSE plan_B).
- • **Markov Decision Processes (MDPs):** Fully stochastic environments modeled via transition probabilities $P(s' | s, a)$ and reward functions $R(s, a, s')$, solved via Bellman Optimality Equations and Dynamic Programming (Value/Policy Iteration).

5.3 Introduction to Generative AI (LLMs & Diffusion Models)

Definition: Generative AI
Generative AI refers to deep learning foundation models trained on massive internet-scale multimodal datasets that can synthesize novel high-fidelity text, images, computer code, audio, and 3D assets from natural language prompts.

1. Large Language Models (LLMs - GPT Series)

- • **Transformer Architecture (Vaswani et al., 2017):** Replaced recurrent networks with the Self-Attention Mechanism, enabling parallel processing of global sequence dependencies:

$$Attention(Q, K, V) = softmax((Q \times K^T) / \sqrt{d_k}) \times V$$

- • **Generative Pre-trained Transformer (GPT):** Autoregressive decoder models trained on Next-Token Prediction on trillions of tokens, fine-tuned via RLHF (Reinforcement Learning from Human Feedback) for alignment and instruction following.

2. Generative Diffusion Models (DALL-E, Stable Diffusion)

- • **Mechanism:**

– • **Forward Diffusion:** Gradually corrupts training images with Gaussian noise over T timesteps until pure noise is obtained.

– • **Reverse Denoising:** Trains a U-Net neural network guided by text embeddings (CLIP) to iteratively predict and subtract noise, generating crystal-clear photorealistic images from pure noise.

5.4 Explainable AI (XAI): Interpretability & Methods

Definition: Explainable AI (XAI)
Explainable AI encompasses the suite of machine learning techniques and tools that enable human stakeholders to understand, interpret, debug, and trust the decisions, internal representations, and predictions made by complex black-box artificial intelligence systems.

XAI Method	Mathematical / Algorithmic Concept	Core Use Case & Strengths
LIME (Local Interpretable Model-agnostic Explanations)	Perturbs input samples around a specific instance, observes black-box prediction changes, and fits a simple interpretable surrogate model (e.g., linear regression/decision tree) locally.	Local explanations for any arbitrary classifier (tabular, text, images). Model-agnostic.
SHAP (SHapley Additive exPlanations)	Derived from cooperative game theory (Lloyd Shapley). Computes the fair marginal contribution of each input feature across all possible feature subsets.	Theoretically grounded, satisfies local accuracy and consistency properties. Standard for tabular data.
Grad-CAM (Gradient-weighted Class Activation Mapping)	Uses gradients of the target class score flowing into the final convolutional layer of a CNN to produce a 2D heat map of prominent visual regions.	Computer vision model debugging (verifying if model looks at tumor or background artifact in X-ray).

5.5 Federated Learning & Edge AI

1. Federated Learning (Privacy-Preserving Decentralized AI)

Definition: Federated Learning (McMahan et al., Google 2017)

Federated Learning is a decentralized machine learning paradigm where thousands of edge client devices (smartphones, IoT sensors, hospital nodes) collaboratively train a shared global model while keeping all raw private training data stored strictly locally on the devices.

- **• The 4-Step FedAvg (Federated Averaging) Cycle:**
  - **1. Server Distribution:** Central server distributes current global model weights w_t to a selected subset of edge clients.
  - **2. Local Training:** Each client trains the model on its private local dataset using Stochastic Gradient Descent (SGD), generating local weights w_t^k.
  - **3. Secure Model Aggregation:** Clients transmit encrypted parameter updates (Δw) back to the server (no raw user data is ever transmitted).
  - **4. Global Model Averaging:** Server computes the weighted average of client weights:

$$w_{(t+1)} = \sum_{k=1}^{K} (n_k / n) \times w_t^k$$

2. Edge AI (On-Device Inference Optimization)

Edge AI executes AI inference directly on local hardware chips (Apple Neural Engine, Qualcomm NPU, Raspberry Pi) without communicating with cloud data centers:

- **• Model Compression Techniques:**
  - **• Quantization:** Reduces numerical precision from 32-bit floating point (FP32) to 8-bit or 4-bit integers (INT8/INT4), slashing memory footprint by 75% and speeding up matrix multiplication on NPUs.
  - **• Network Pruning:** Identifies and removes redundant weights or entire attention heads that contribute minimally to model accuracy.
  - **• Knowledge Distillation:** Trains a compact 'Student' neural network to mimic the soft probability distributions of a massive 'Teacher' model.

5.6 High-Yield University Exam Question Bank & Model Answers

Part A: Short Answer Questions (2–3 Marks)

- **Q1. What is PDDL in Automated Planning? What are its components?**

Answer: PDDL (Planning Domain Definition Language) is a standardized formal language for classical planning problems. It specifies: (1) Objects, (2) Initial State (conjunction of ground literals), (3) Goal State, and (4) Actions (defined by Parameters, Preconditions, and Effects: Add/Delete lists).

- **Q2. Explain the Self-Attention mechanism in Transformer models.**

Answer: Self-Attention computes dynamic contextual relationships between all words in a sequence simultaneously using Query (Q), Key (K), and Value (V) projections: $\text{Attention}(Q, K, V) = \text{softmax}((Q K^T) / \sqrt{d_k}) V$, eliminating recurrent bottlenecks.

- **Q3. Differentiate between LIME and SHAP in Explainable AI.**

Answer: LIME is a local surrogate method that perturbs input data to approximate a black-box model locally with an interpretable linear model. SHAP uses game-theoretic Shapley values to calculate the exact, fair additive marginal contribution of each feature across all subsets.

- **Q4. How does Federated Learning protect user data privacy?**

Answer: Federated Learning trains models directly on local user devices. Only encrypted model weight updates (Δw) are sent to the central server for global aggregation via FedAvg; raw user training data never leaves the device.

Part B: Long Answer Questions (8–10 Marks)

- **Q5. Explain Classical Planning in AI. Detail STRIPS/PDDL schemas, Forward State-Space Search, Backward Goal Regression, and Planning Graphs.**

Model Answer Outline:

- 1. Definition of Classical Planning: State representations, open-world vs closed-world assumption.
- 2. PDDL Action schemas: Blocks World or Air Cargo Transport example showing Preconditions and Add/Delete Effects.
- 3. Forward Search (Progression) vs Backward Search (Regression): Algorithms, branching factors, and applicability.
- 4. Planning Graphs (Graphplan): Proposition levels, Action levels, Mutual Exclusion (Mutex) relations (Inconsistent effects, Interference, Competing needs).
- 5. Heuristics extraction from planning graphs for A* planning search.

- **Q6. Discuss Emerging Frontiers in AI: Generative AI (Transformers/Diffusion), Explainable AI (XAI), and Edge AI with Model Compression.**

Model Answer Outline:

- 1. Generative AI: Architecture of Transformers (Self-attention, Multi-head attention, FFN) and Generative Diffusion Models (Forward noise addition and reverse U-Net denoising).
- 2. Explainable AI (XAI): Need for interpretability in healthcare/finance; LIME local surrogates, SHAP Shapley values, and Grad-CAM visual heatmaps.
- 3. Federated Learning: Architectural pipeline diagram, FedAvg formula, and differential privacy guarantees.

- 4. Edge AI: On-device deployment constraints, Quantization (FP32 → INT8), Weight Pruning, and Knowledge Distillation (Teacher-Student).

5.7 Unit V Summary & Quick Revision Sheet

Planning & Modern AI Domain	Key Theoretical & Technical Takeaways for Exams
Classical Planning	PDDL / STRIPS representation (Preconditions & Effects). Progression vs Regression. Graphplan & Mutex relations.
Hierarchical & Stochastic	HTN (Task decomposition). Conformant (No sensors), Contingent (With sensing branches), MDPs (Bellman equations).
Generative AI	Transformers (Self-attention Attention(Q,K,V)=softmax(QK^T/√d)V), GPT (Autoregressive), Diffusion (Reverse denoising).
Explainable AI (XAI)	LIME (Local linear surrogate), SHAP (Game-theoretic Shapley values), Grad-CAM (CNN gradient heatmaps).
Federated Learning & Edge AI	FedAvg decentralization (w_(t+1) = ∑(n_k/n)w_k). Edge AI: Quantization (INT8), Pruning, Knowledge Distillation.