Unit 6

Emerging Trends in AI

(2 Hours)

6.1 Generative AI

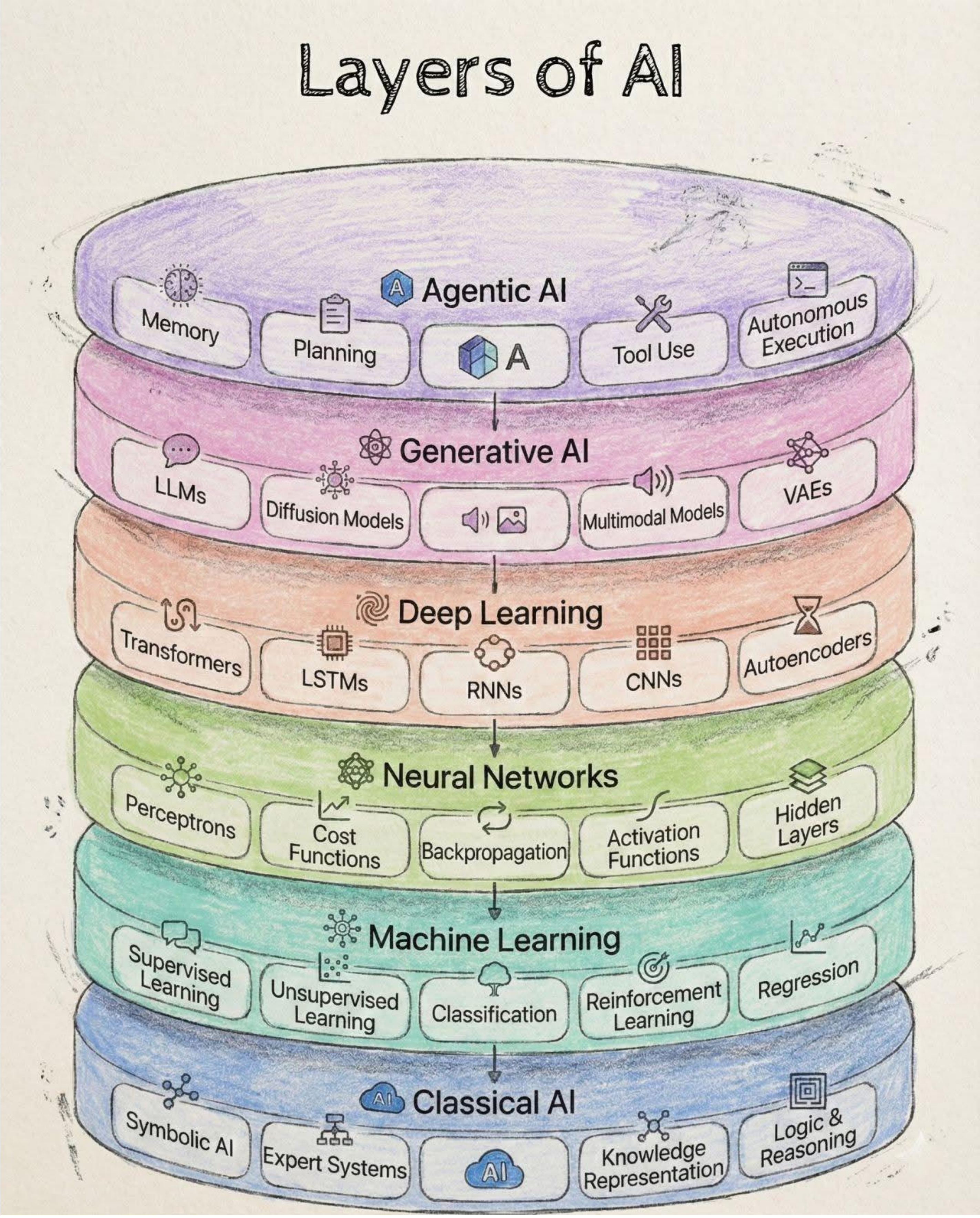
6.2 Explainable AI

6.3 Ethical AI

6.4 Multi-model AI

6.5 Integration of AI in different sectors: Health care, Cyber Security, IOT, Quantum Computing

Sanjeev Thapa, Er. DevOps, SRE, CKA, RHCSA, RHCE, RHCSA-Openstack, MTCNA, MTCTCE, UBSRS, HEv6, Research Evangelist



Sanjeev Thapa, Er. DevOps, SRE, CKA, RHCSA, RHCE, RHCSA-Openstack, MTCNA, MTCTCE, UBSRS, HEv6, Research Evangelist

Based on the image structured notes breaking down the "Layers of AI." The diagram illustrates a hierarchy where each advanced layer is built upon the foundational concepts

Level 1: Classical AI (The Foundation)

The base layer representing the earliest approaches to artificial intelligence, relying on explicit rules and human-defined logic.

- **Symbolic AI:** AI based on high-level symbolic (human-readable) representations of problems.
- **Expert Systems:** Computer systems emulating the decision-making ability of a human expert.
- **Knowledge Representation:** How information is stored so an AI can utilize it.
- **Logic & Reasoning:** Using formal logic to deduce new facts from existing data.

Level 2: Machine Learning

Moving beyond strict rules, this layer introduces systems that can learn patterns from data.

- **Supervised Learning:** Learning from labeled training data (e.g., input-output pairs).
- **Unsupervised Learning:** Finding hidden patterns in unlabeled data.
- **Reinforcement Learning:** Learning through trial and error (rewards and punishments).
- **Classification:** Categorizing data into classes (e.g., "spam" vs. "not spam").
- **Regression:** Predicting continuous values (e.g., housing prices).

Level 3: Neural Networks

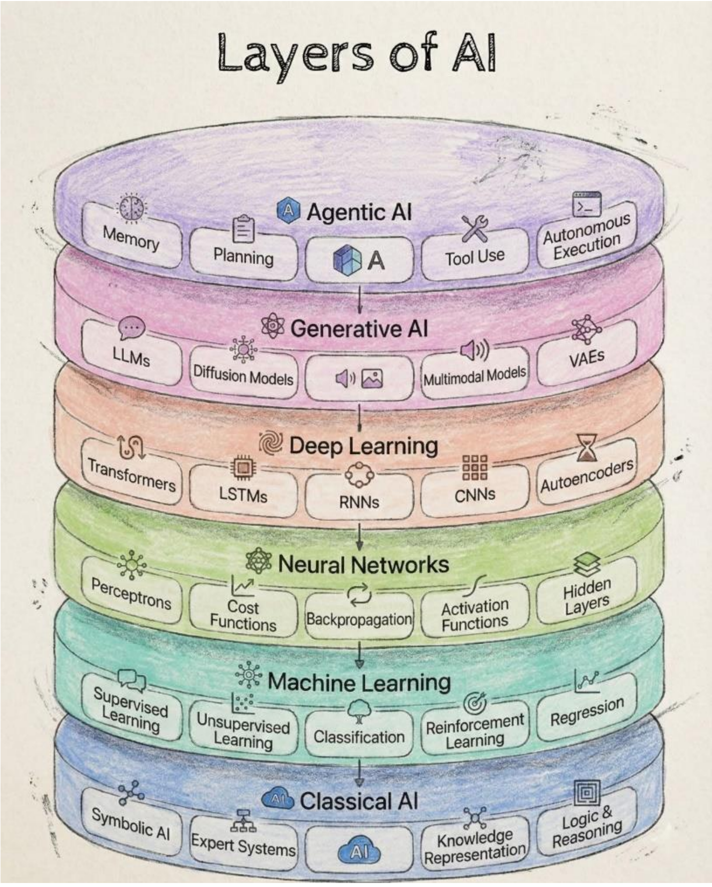
Algorithms inspired by the human brain's structure, designed to recognize relationships in data.

- **Perceptrons:** The simplest type of artificial neural network (a single neuron).
- **Hidden Layers:** Layers of neurons between the input and output that process features.
- **Backpropagation:** The method used to calculate errors and update the network to improve accuracy.
- **Activation Functions:** Mathematical equations that determine if a neuron should "fire."
- **Cost Functions:** Measures how wrong the model's predictions are (used to guide learning).

Level 4: Deep Learning

A subset of machine learning using multi-layered (deep) neural networks to solve complex problems.

- **Transformers:** The architecture behind modern LLMs (focuses on attention mechanisms).
- **CNNs (Convolutional Neural Networks):** Specialized for processing grid data like images.
- **RNNs (Recurrent Neural Networks):** Designed for sequential data (like time series or text).
- **LSTMs (Long Short-Term Memory):** A type of RNN capable of learning long-term dependencies.
- **Autoencoders:** Neural networks used for data compression and strictly learning efficient codings.



Level 5: Generative AI

AI capable of creating new content (text, images, audio) rather than just analyzing existing data.

- **LLMs (Large Language Models):** Models trained on vast text to understand and generate human language.
- **Diffusion Models:** Used primarily for generating high-quality images.
- **Multimodal Models:** AI that can process multiple types of inputs simultaneously (e.g., text and images).
- **VAEs (Variational Autoencoders):** A generative model often used in image generation.

Level 6: Agentic AI (The Top Layer)

The current cutting edge: AI systems that can act autonomously to achieve goals, not just generate text.

- **Planning:** The ability to break a complex goal into a sequence of actionable steps.
- **Memory:** Retaining information over time to maintain context.
- **Tool Use:** The ability to call external software (calculators, web browsers, APIs) to perform tasks.
- **Autonomous Execution:** Carrying out tasks with minimal human intervention.

6.1 Generative AI (GenAI)

Moving from "analyzing" data to "creating" data.

- **Definition:** A type of AI that can create new content (text, images, audio, video, code) in response to user prompts. Unlike traditional AI that classifies data (e.g., "Is this a cat?"), GenAI creates data (e.g., "Draw a cat in the style of Van Gogh").
- **How it Works (Simplified):** It learns patterns from massive datasets and uses probability to predict the "next piece" of information.
  - **Text:** Uses **Transformers** (like GPT) to predict the next word.
  - **Images:** Uses **Diffusion Models** (adds noise to an image until it is static, then learns to reverse the process to create a clear image from static).
- **Key Examples:**
  - **Text:** ChatGPT (OpenAI), Claude (Anthropic), Llama (Meta).
  - **Image:** Midjourney, Stable Diffusion, DALL-E 3.
  - **Code:** GitHub Copilot, Code Llama.
- **Teaching Tip:** *Ask students to compare Google Search (Retrieval) vs. ChatGPT (Generation). Google finds what exists; ChatGPT creates what might not exist yet.*

6.2 Explainable AI (XAI)

Solving the "Black Box" problem.

- **Definition:** A set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms.
- **The Problem (Black Box):** Deep Learning models (like Neural Networks) are often so complex that even their creators don't know *why* they made a specific decision.
- **The Solution (Glass Box):** XAI attempts to show the "reasoning" behind the output.
- **Key Concepts:**
  - **Interpretability:** Can we understand the cause and effect?
  - **Transparency:** Is the model architecture known?
  - **Feature Importance:** Highlighting which parts of the input led to the decision (e.g., "The AI denied the loan because of 'Debt-to-Income Ratio', not 'Zip Code'").
- **Real-World Example:**

  - **Healthcare:** If an AI predicts a patient has cancer, XAI highlights the exact pixels on the X-ray that triggered the diagnosis so a doctor can verify it.
  - **Finance:** Explaining why a credit card transaction was flagged as fraud.

6.3 Ethical AI

Ensuring AI helps rather than harms.

- **Definition:** The practice of designing and deploying AI systems that align with moral principles and societal values.
- **Core Issues:**
  - **Bias & Fairness:** AI models can inherit racism or sexism from their training data (e.g., facial recognition working poorly on darker skin tones).
  - **Privacy:** How is user data used? (e.g., Should Zoom use your calls to train its AI?).
  - **Accountability:** Who is responsible if an AI car crashes? The driver, the programmer, or the manufacturer?
  - **Deepfakes:** The malicious use of AI to create fake videos/audio of real people.
- **Mitigation:**
  - **RLHF (Reinforcement Learning from Human Feedback):** Humans rate AI responses to teach it "good" behavior.
  - **Red Teaming:** Hiring hackers to try and break the AI to find safety flaws before release.

6.4 Multi-modal AI

AI that has "eyes" and "ears," not just "text."

- **Definition:** AI models that can process and understand multiple types of input (modalities) simultaneously—such as text, images, audio, and video—rather than just one.
- **Evolution:**
  - *Unimodal:* An AI that only does text (old ChatGPT).
  - *Multimodal:* An AI that you can talk to, show a picture of your fridge, and ask for a recipe (GPT-4o, Gemini 1.5).
- **How it works:** It maps different data types into a shared "vector space." The concept of a "dog" relates to the word "dog," the sound of a "bark," and the image of a "puppy."

- **Examples:**
  - **GPT-4o / Gemini:** Can see images and hear audio in real-time.
  - **CLIP (OpenAI):** Connects text to images (used to categorize photos without labels).
  - **ImageBind (Meta):** Connects text, audio, visual, thermal, and depth data.

6.5 Integration of AI in Different Sectors

Sector	**Key Applications**	**Example**
Healthcare	**Drug Discovery:** Predicting protein folding to find cures faster. **Diagnostics:** Analyzing CT scans/MRIs faster than radiologists.	**AlphaFold (Google):** Solved a 50-year-old biology problem by predicting 3D protein structures.
Cyber Security	**Anomaly Detection:** AI monitors network traffic 24/7 to find weird patterns that indicate a hack. **Phishing Defense:** Analyzing email context to catch scams that bypass traditional filters.	**Darktrace:** Uses AI to fight cyber-attacks in real-time ("AI vs. AI" warfare).
IoT (AIoT)	**Edge AI:** Running AI on the device (camera, fridge) rather than the cloud to save bandwidth and improve privacy. **Predictive Maintenance:** Sensors listen to factory machine vibrations to predict when they will break *before* they fail.	**Smart Cities:** Traffic lights that change timing based on real-time traffic flow (cameras + AI).

Sector	**Key Applications**	**Example**
Quantum Computing	**Quantum Machine Learning (QML):** Using quantum computers to process massive AI datasets exponentially faster. **Optimization:** Solving complex logistics problems that are impossible for classic supercomputers.	**Battery Design:** Using Quantum AI to simulate molecular interactions to invent better EV batteries.