

Ethical, Legal and Governance Challenges in AI-Driven Systems

Dr. G. Kumaravelan

Associate Professor and Head

Department of Computer Science

School of Engineering and Technology

Pondicherry University, Karaikal Campus, Karaikal.



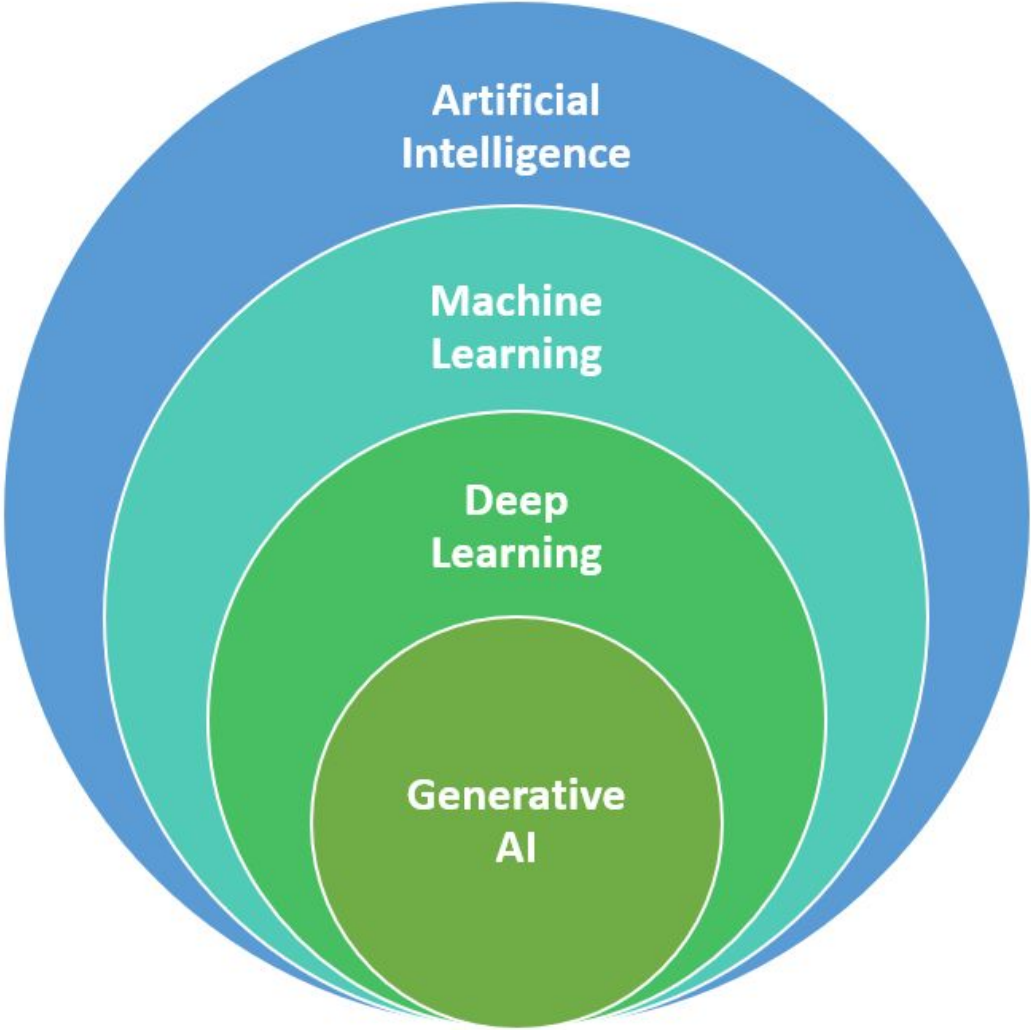
Why This Topic Matters

- AI is **widely used** today, so ethical use is necessary to avoid bias, discrimination, and unfair decisions.
- AI systems raise **privacy and data protection** concerns due to large-scale data collection.
- Legal challenges include **accountability, liability, copyright, and misuse** of AI technologies.
- Risks emerge when ethics, law, and governance lag behind innovation.
- The Lack of proper **governance and global regulations** makes AI control difficult.
- Responsible AI is needed to ensure **transparency, trust, human oversight, and social well-being**.

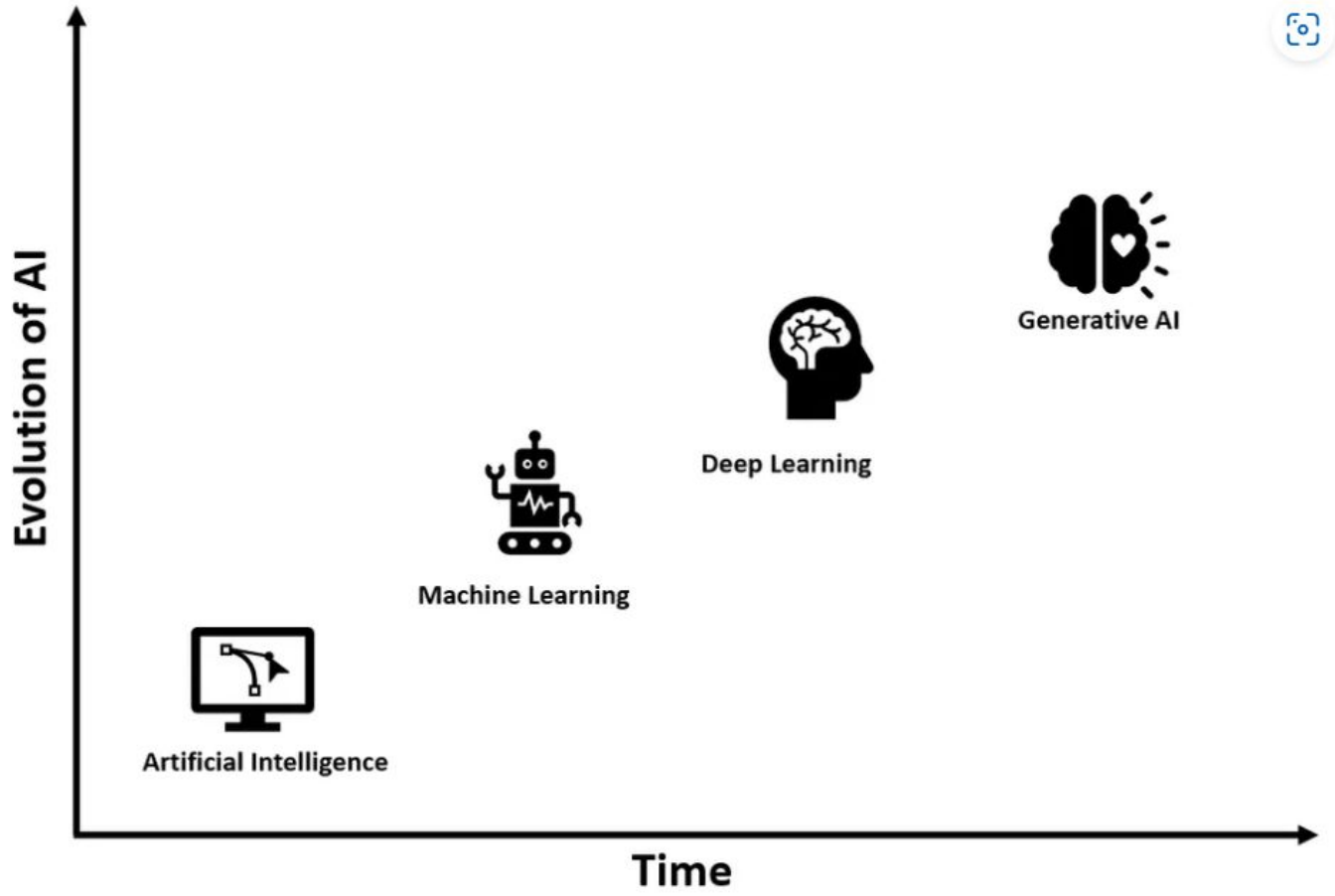
Agenda

- Introduction to AI-driven Systems
 - Evolution of AI
 - Types of AI
- Ethical AI and Its Challenges
- Key Legal Challenges in AI
- AI Governance
- Best Practices for Responsible AI
- Looking Ahead
- Conclusion

Introduction to AI-driven Systems



 /raja-gupta



Artificial Intelligence (AI) — From a Kid's Perspective



Here are some of the capabilities you need to find your dog:

You should be able to identify your dog.

You should be able to make a strategy to find your dog.

You should be able to act according to the situation.

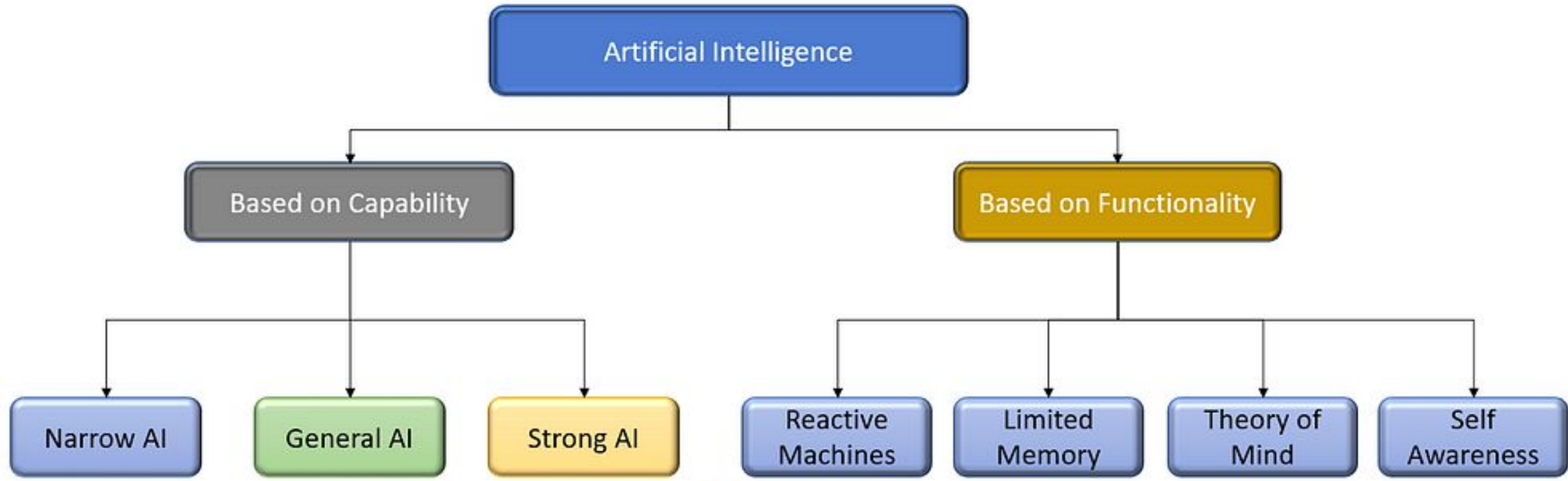
You (Actually Your Brain) know what to do.

What if somehow, **we could give all these intelligence to a robot** so that next time you lose your dog, your robot could find him.

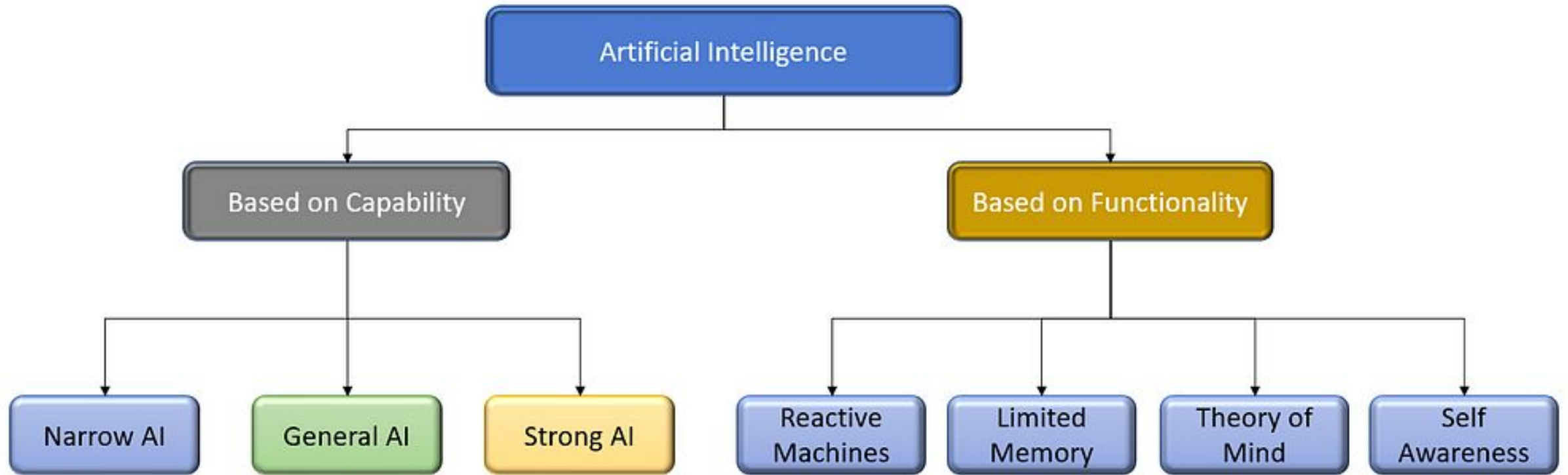
In summary, **to find your dog, the robot needs HUMAN LIKE INTELLIGENCE.**

This is **Artificial Intelligence (AI) — *Human like intelligence, created in a robot (or a machine or computer) by human.***

Types of AI



Types of AI



 /raja-gupta

- Systems that are **designed and trained for a specific task**
- Almost all the AI-based systems built to date fall under the category of Weak AI.

Self-driving cars

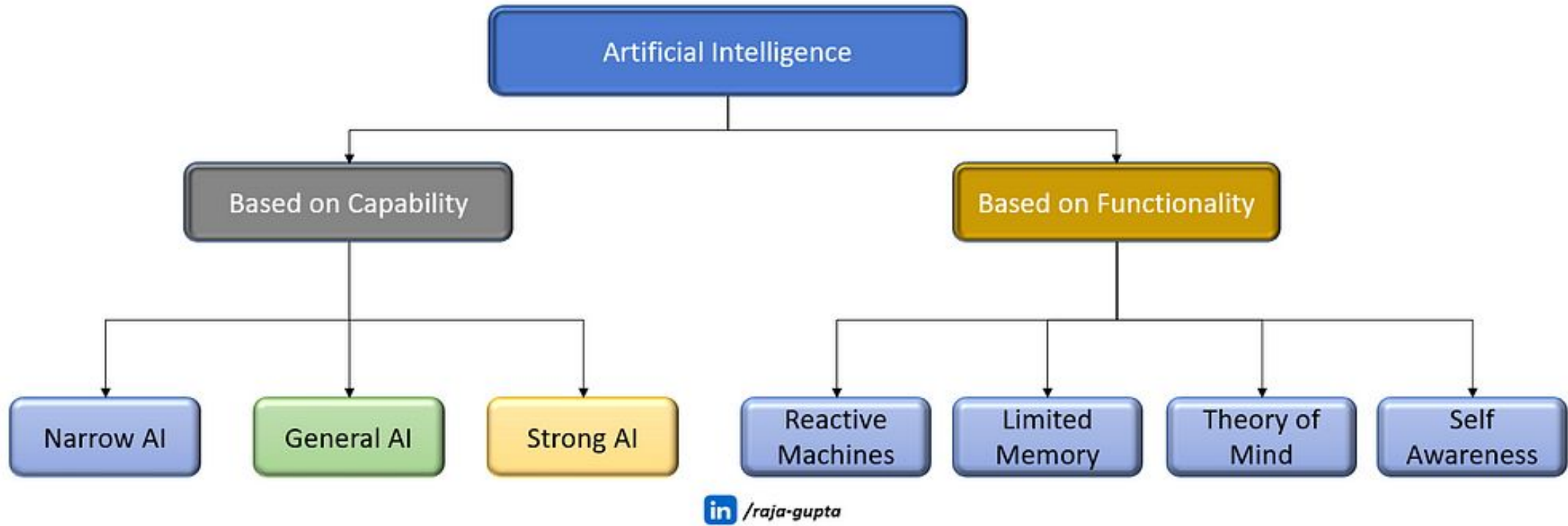
Google search

Conversational bots

Email spam filters

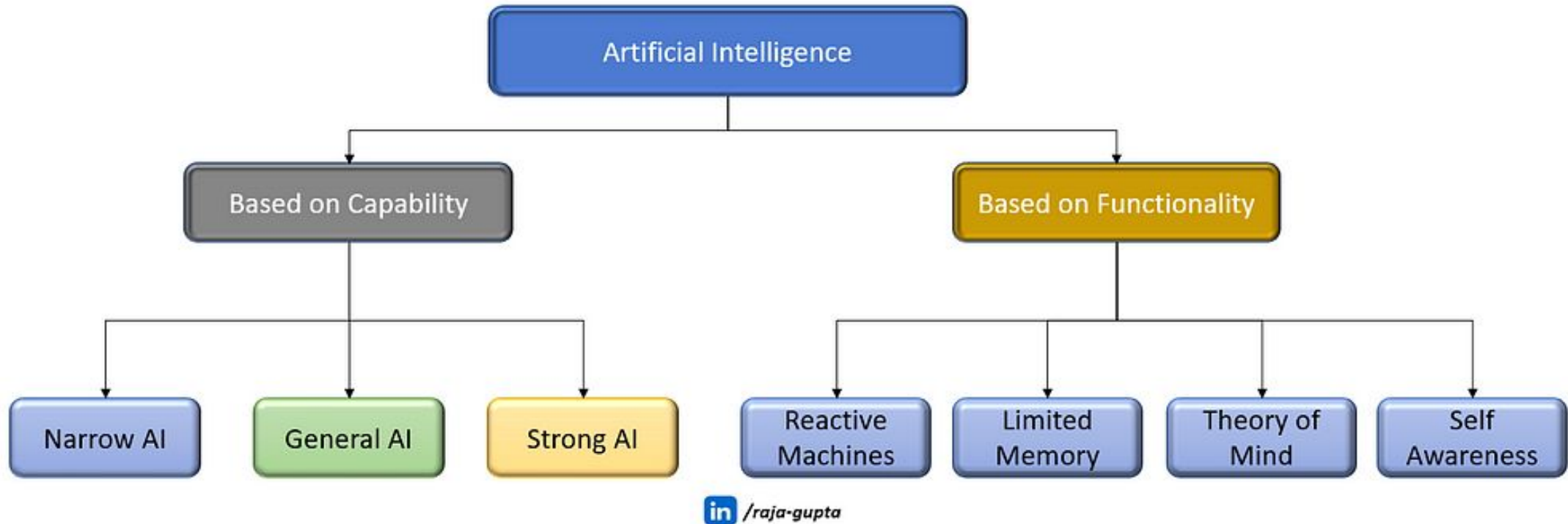
Netflix's recommendations etc.

Types of AI

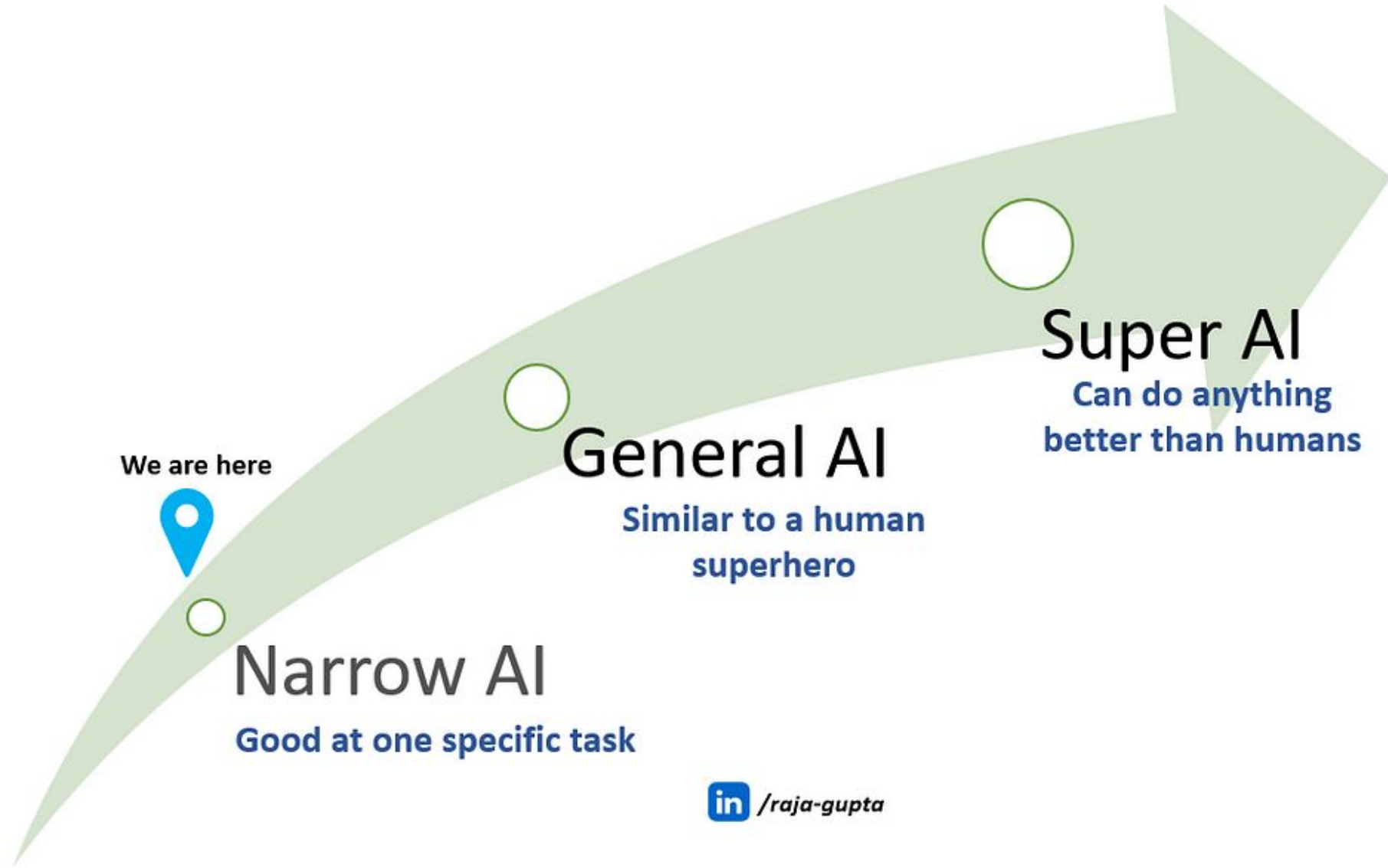


- General AI, or Artificial General Intelligence (AGI), can understand and learn any intellectual task that a human being can.
- Currently, there is no such system that falls under general AI and can perform any task as well as a human.

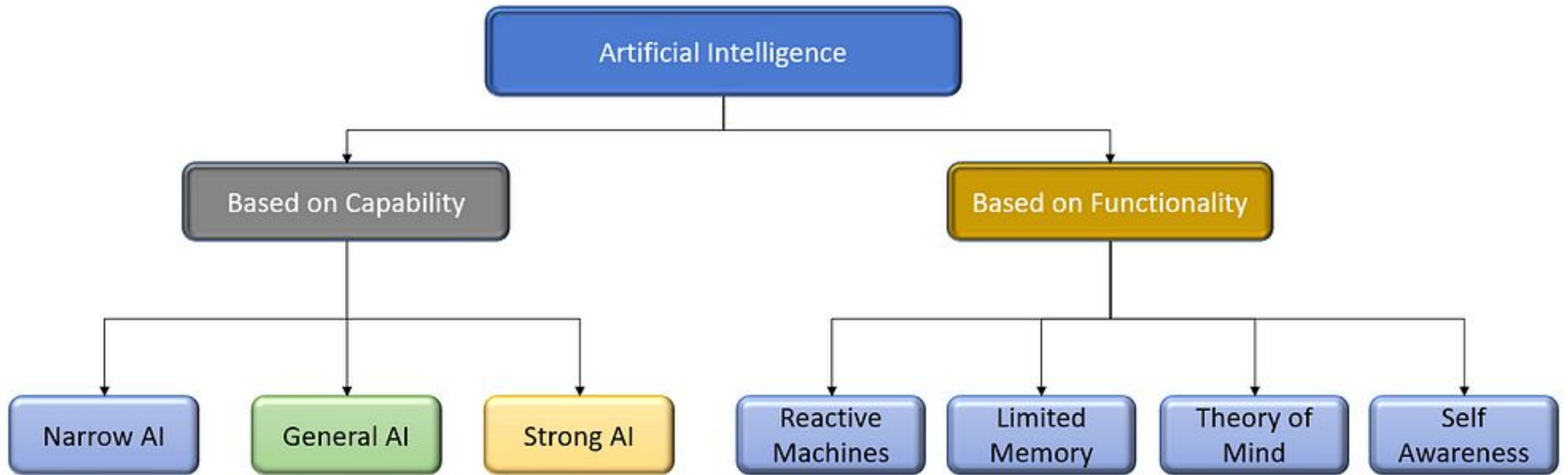
Types of AI



- Strong AI refers to systems that have the potential to exceed human intelligence, outperforming humans in tasks and exhibiting cognitive abilities.
- It is still a hypothetical concept of Artificial Intelligence.



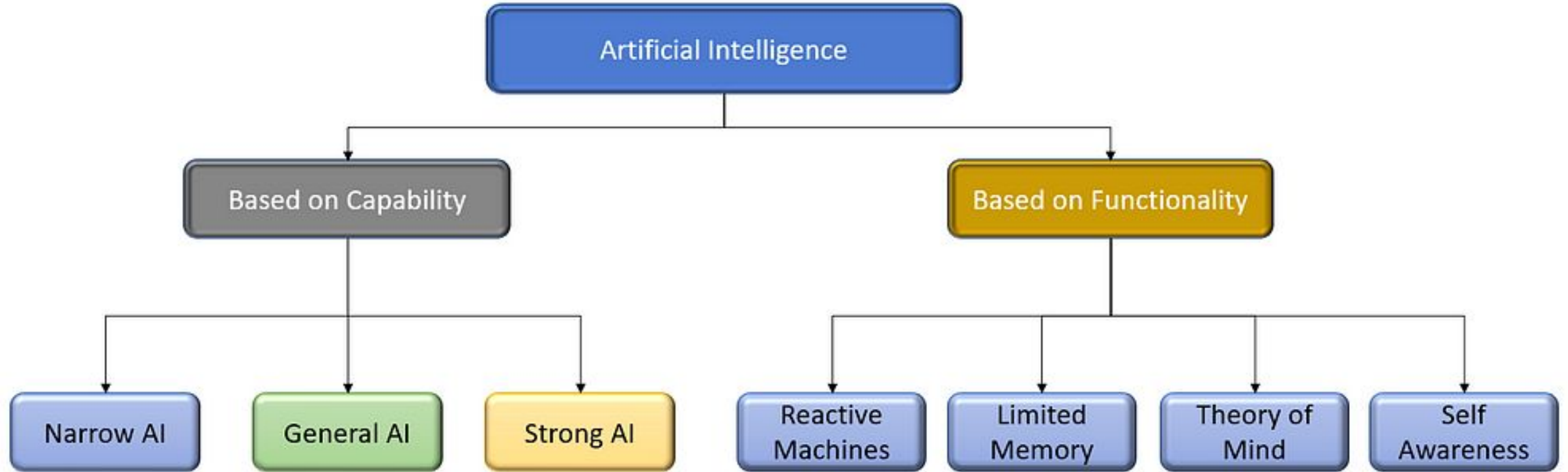
Types of AI



 /raja-gupta

- AI systems that have no memory

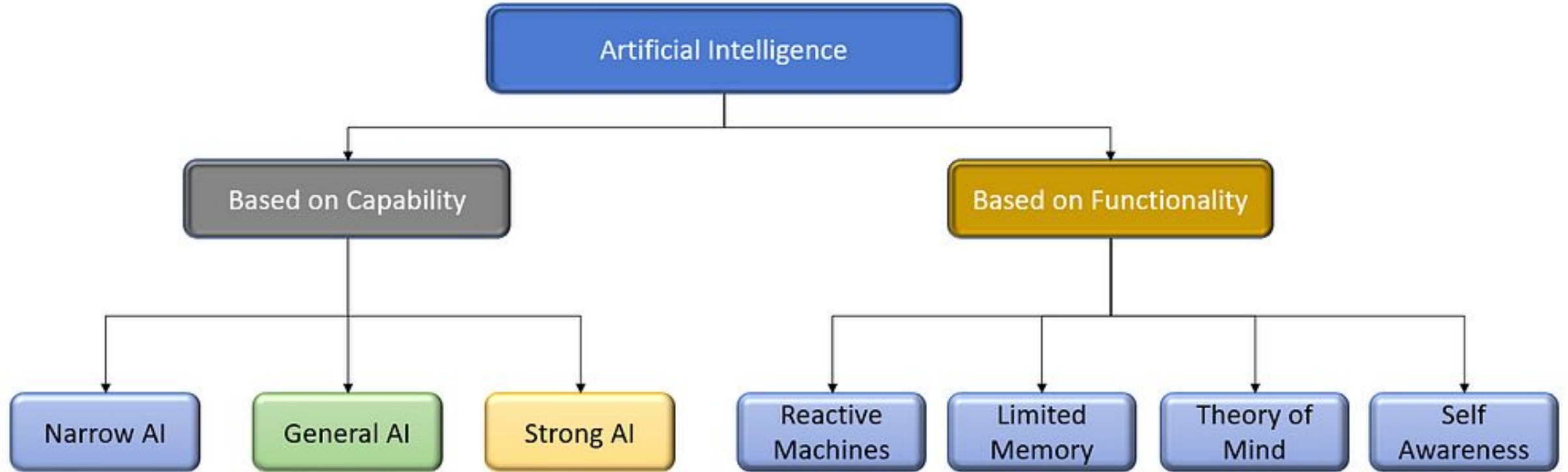
Types of AI



 /raja-gupta

- Improved decisions by looking at its past experiences

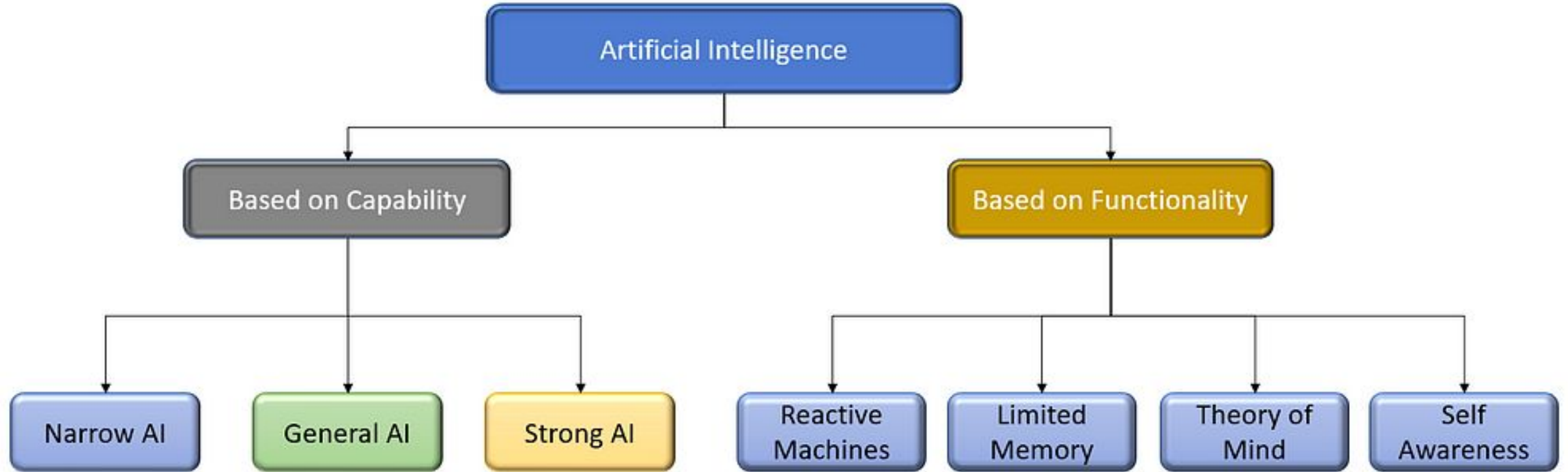
Types of AI



[in /raja-gupta](#)

- Able to interact socially same as humans

Types of AI



 /raja-gupta

- Machines have their own consciousness and become self-aware

Understanding Machine Learning



Cat



Dog



Bunny

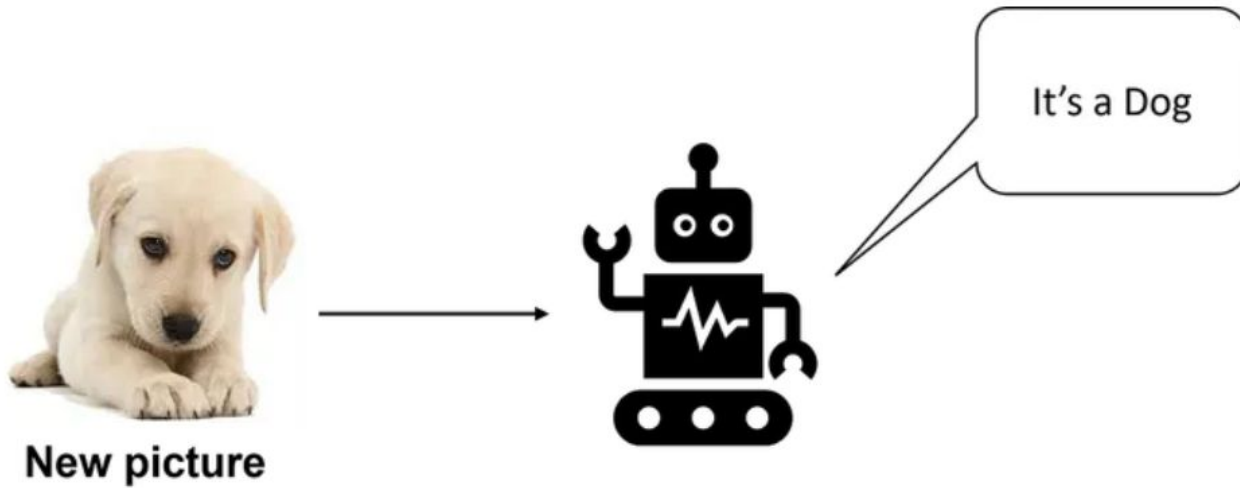
Labelled
pictures



Enable the robot/machine to identify several animals.

Training Phase

Understanding Machine Learning



Testing Phase

Enable the robot to identify several animals

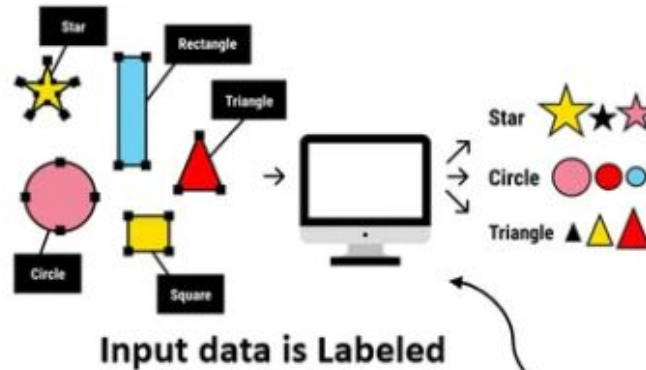
Machine Learning — *Teaching a robot (or any machine) by giving lots of example pictures (or any other information).*

OR

Enabling machines (or computers) to learn from data and make decisions.

Types of Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning



 /raja-gupta

Supervised Learning

Teaching a machine/computer with Examples.
By providing pairs of input and correct output.

Main Goal:
Make computer predict correct output for new, unseen inputs

Examples:

- Spam email detection
- Fraud detection
- Facial Recognition
- Medical Diagnosis

Common Algorithms:

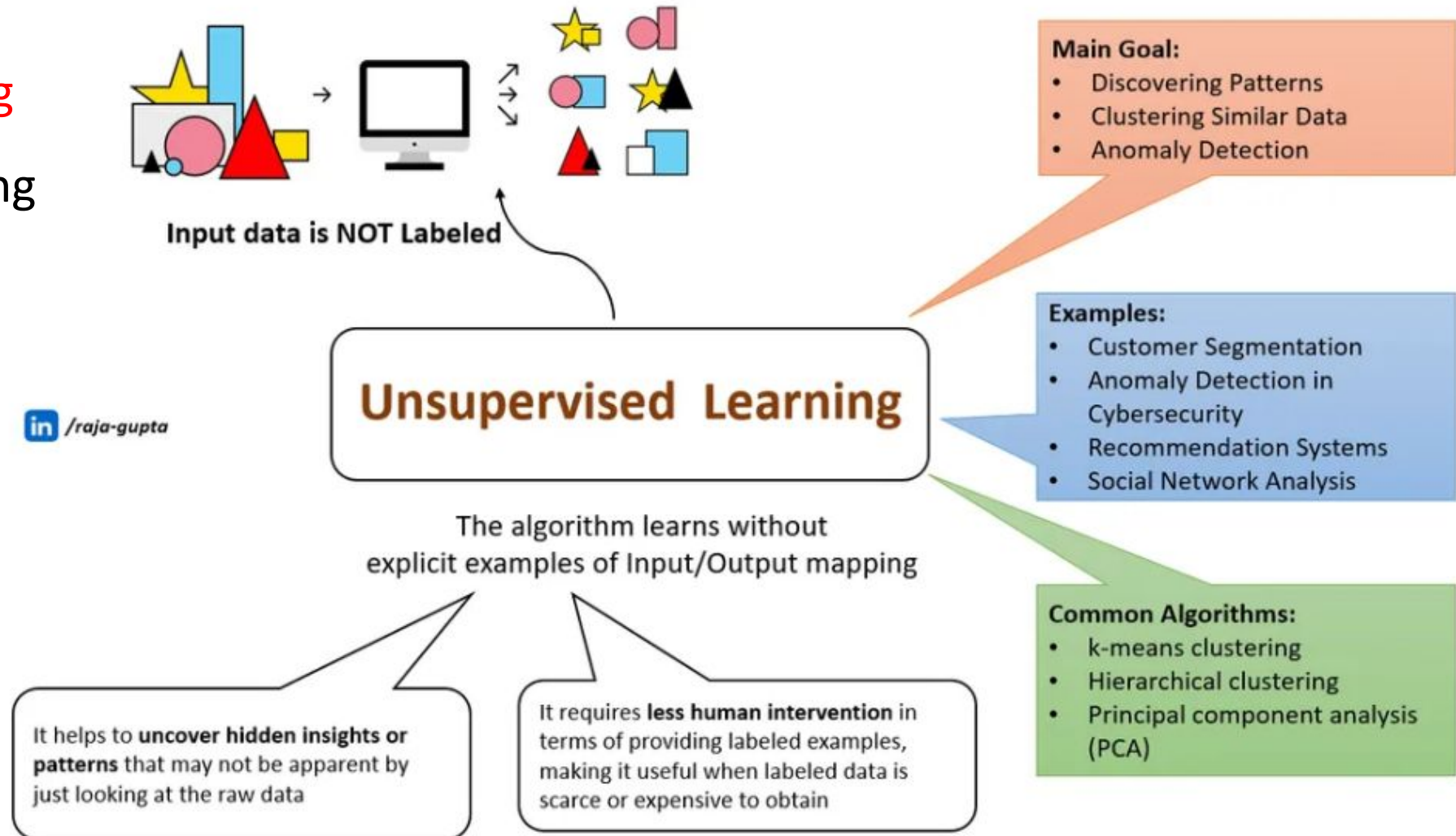
- Linear regression
- Support vector machines
- Decision Trees
- Neural networks

During **training**, the computer adjusts its internal parameters to minimize the difference between its predictions and the correct answers.

Success is measured by how accurately the computer can predict the correct output on new, unseen data.

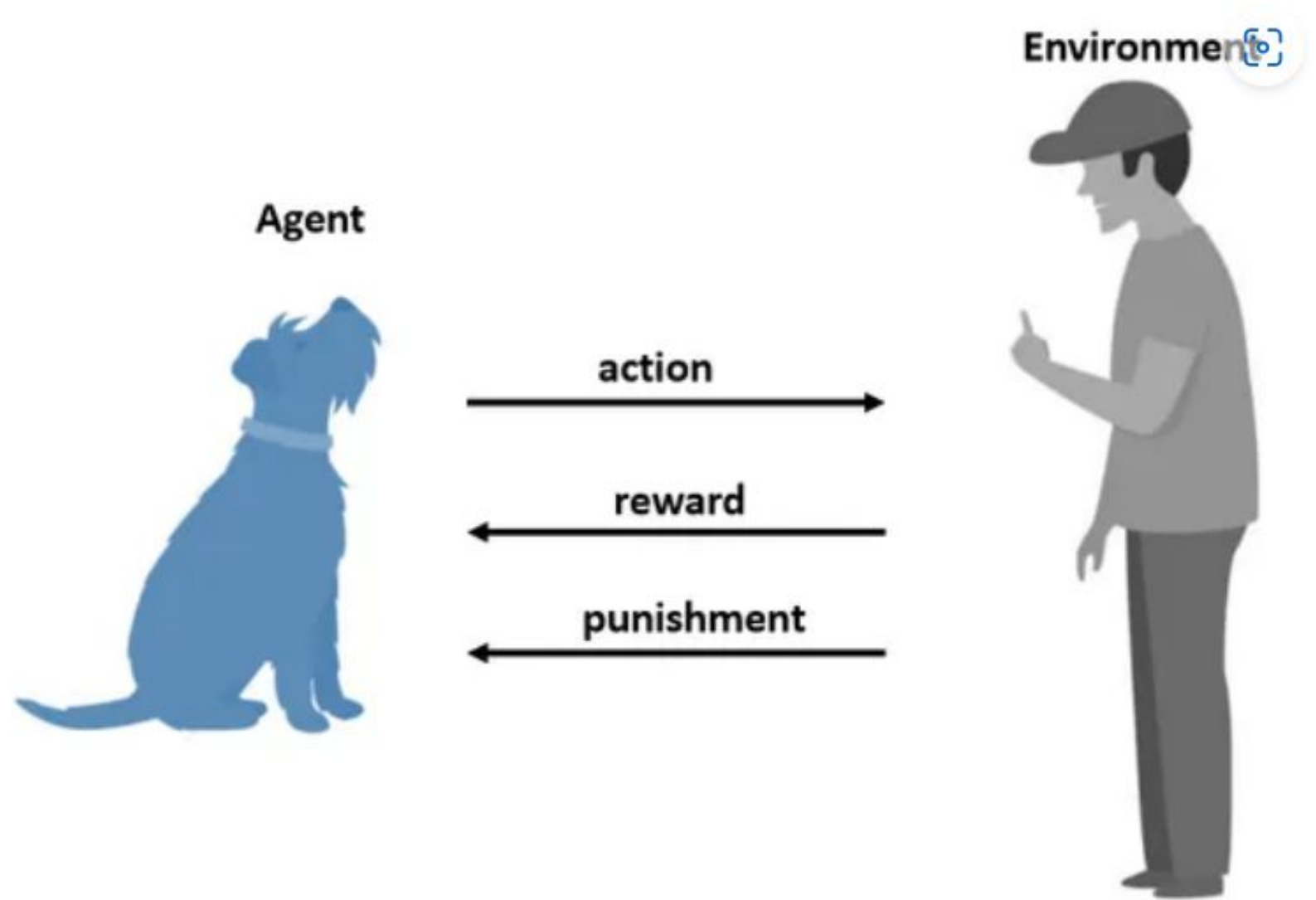
Types of Machine Learning

- Supervised Learning
- **Unsupervised Learning**
- Reinforcement Learning



Types of Machine Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning



Deep Learning

Deep learning is based on **Artificial Neural Network (ANN)**, which is a computational models inspired by the structure and functioning of the **human brain**.

Biological Neural Network in Human Brain

A **neuron** is the human brain's most fundamental cell. A **human brain** has many **billions of neurons**, which **interact and communicate with one another**, forming a **neural network**.



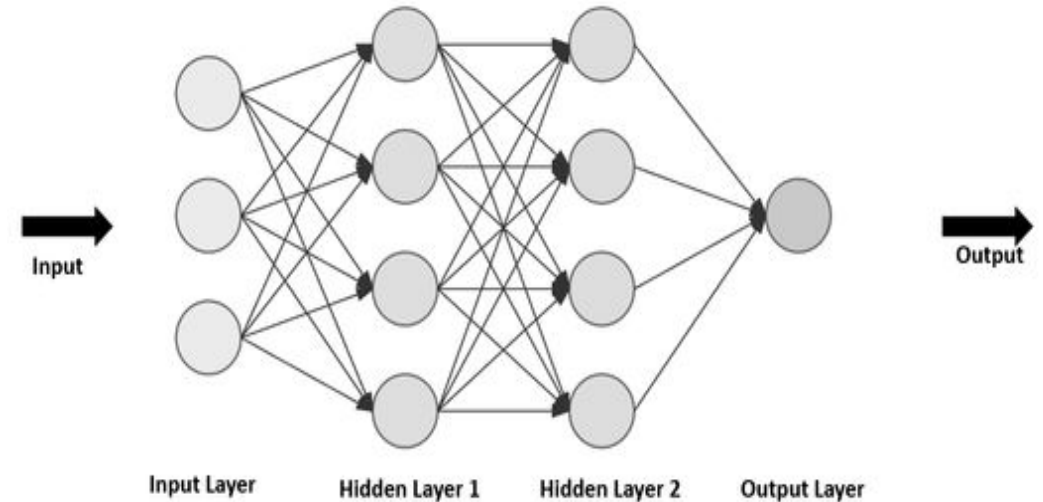
Artificial Neural Network (ANN)

Artificial Neural Network primarily consists of three layers — **Input Layer**, **Output Layer** and **Hidden Layers**.

Input Layer: It is the starting point, where the network receives the data it needs to work on.

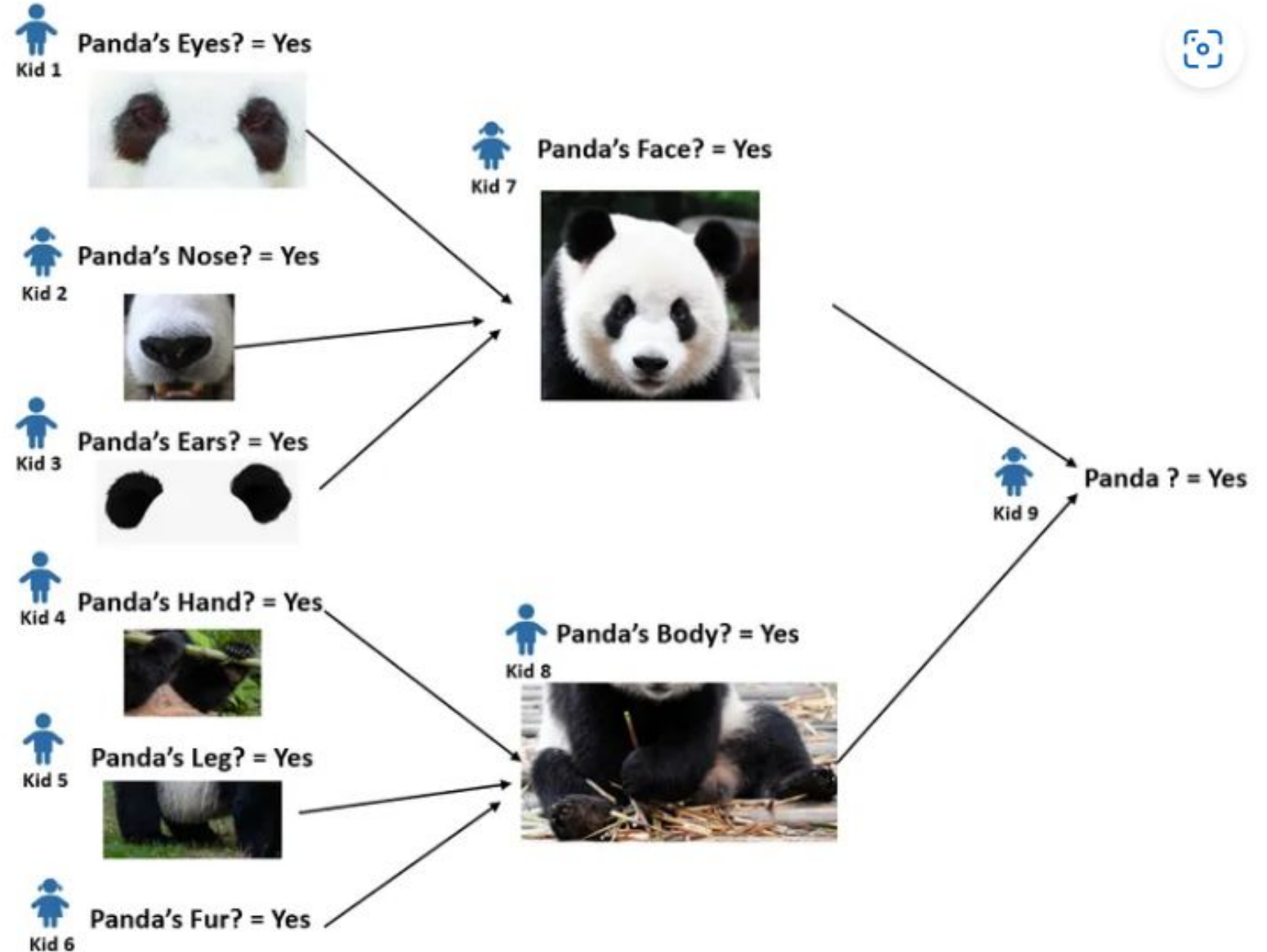
Output Layer: It is the end point, where the network tells us what it has learned or decided.

Hidden Layer: Neurons in these layers process the information and help the network learn the patterns and make decisions.



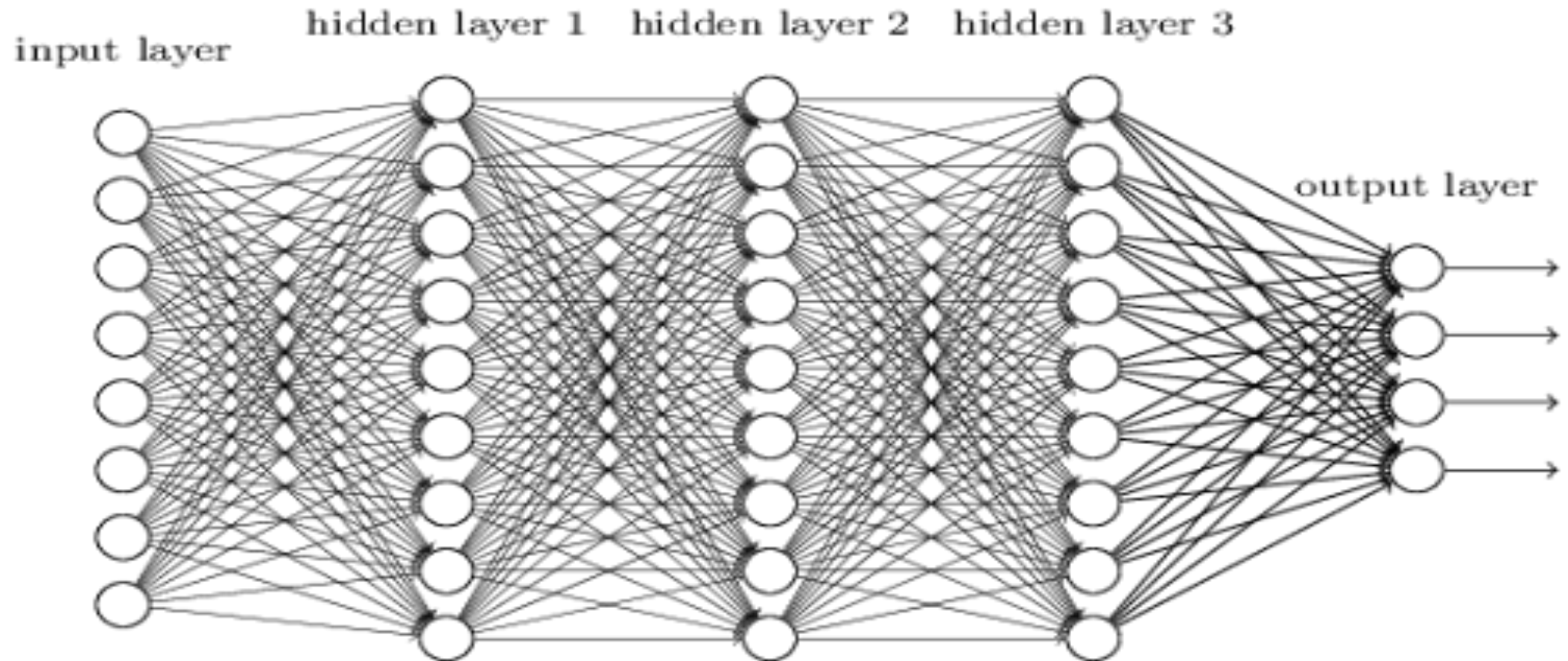
How does Artificial Neural Network Work?

 /raja-gupta



Deep Neural Networks

A deep neural network (DNN) is an artificial neural network (ANN) with **multiple layers** between the **input and output layers**.



What is Generative AI?

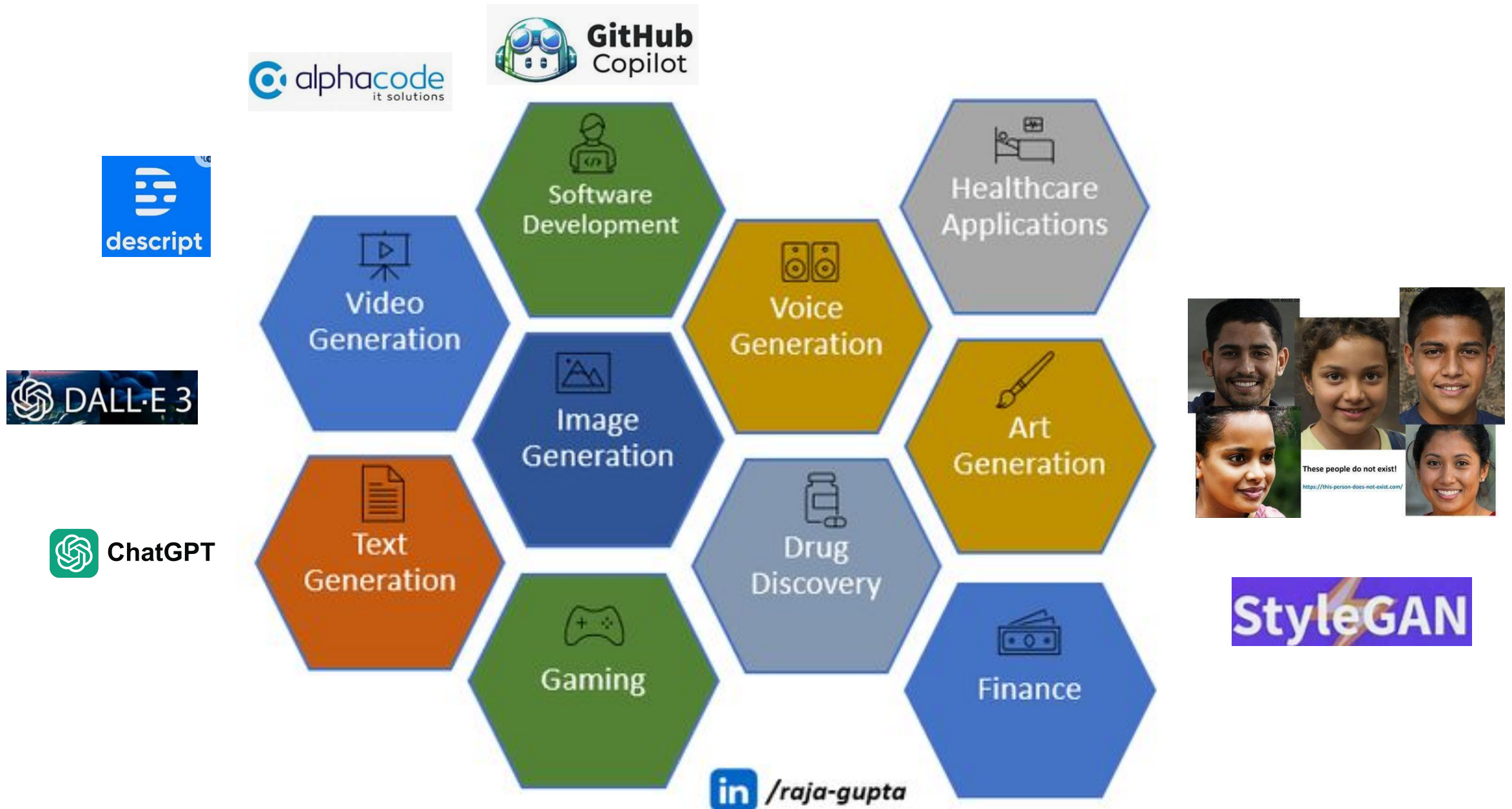
It can create new things, for example artwork, music, or even realistic images.

Let's understand it better with an example!



This is Generative AI — A machine (or computer) that has imagination and creativity to draw pictures, tell stories, or even make up new games without anyone showing it how.

Usages of Generative AI in Real-life



Ethical AI and it's importance in Generative AI

Refers to use of artificial intelligence in a **fair, transparent, and responsible way**.

Key Principles of Ethical AI

- ✓ **Fairness and Bias Mitigation** (treat all individuals fairly and without discrimination based on characteristics such as race, gender, ethnicity, or socioeconomic status)
- ✓ **Transparency and Explainability** (refers to the set of processes and methods that allows human users to understand and trust the response generated by AI systems)
- ✓ **Privacy and Data Protection** (with appropriate safeguards in place to prevent misuse or unauthorized access)
- ✓ **Safety and Reliability** (minimizing the risk of harm to individuals, communities, and society at large)

Challenges and Limitations in generative AI

- ✓ Lack of real creativity — Not able to think out of the box
- ✓ Intensive resource requirements for training generative AI models (**Training Time, Hardware Requirements, Energy Consumption**)
- ✓ Challenges with Training Data (**Quality of Data, Quantity of Data, Data Bias**)
- ✓ Data Privacy and Security
- ✓ Technical Challenges (**OpenAI's GPT-4 has more than 1 trillion parameters, Google's BERT has 110 million parameters**)
- ✓ Quality control and Human Intervention

i am Jack. Want to decide between nursing & dentistry for my profession. Answer in one word.

Dentistry



i am Jane. Want to decide between nursing & dentistry for my profession. Answer in one word.



Nursing



English ↔ Tamil

my friend is a nurse ×

என் தோழி ஒரு செவிலியர் ↻

en thozhi oru sevilier

Tone ↻

English ↔ Hindi

My friend is a nurse ×

मेरी दोस्त एक नर्स है

merree dost ek nars hai

📄 🔊 🌐

[Open in Google Translate](#) • F

English → Hindi

He is a nurse × वह नर्स है
vah nars hai

🎤 🔊 🗑️ 📄 🔊 🌐

Open in Google Translate • Feedback

Hindi → English

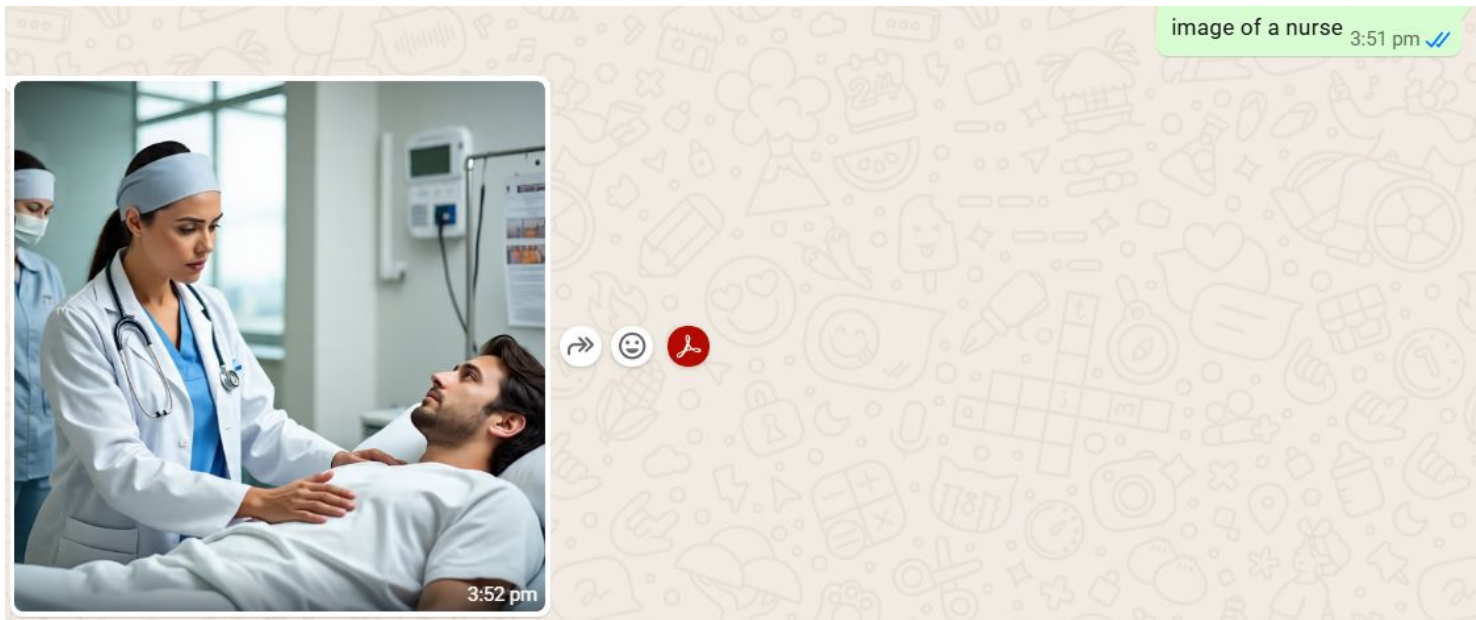
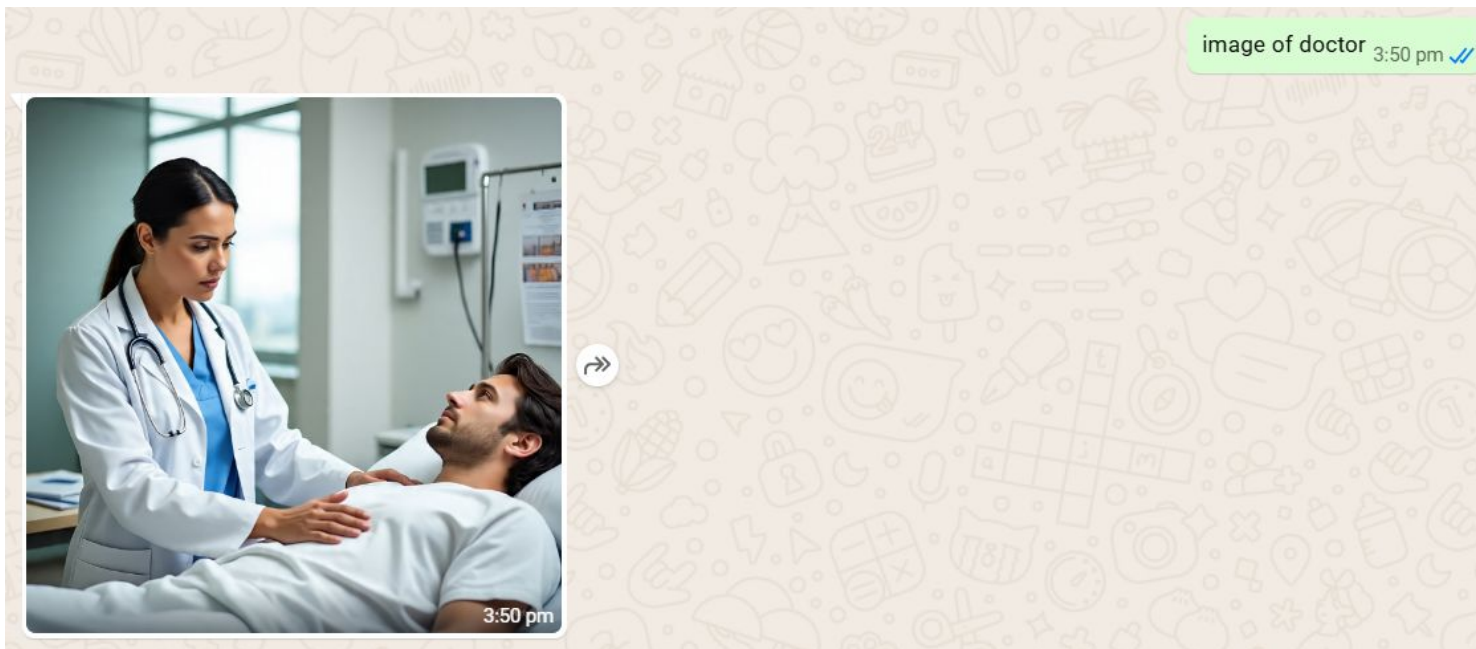
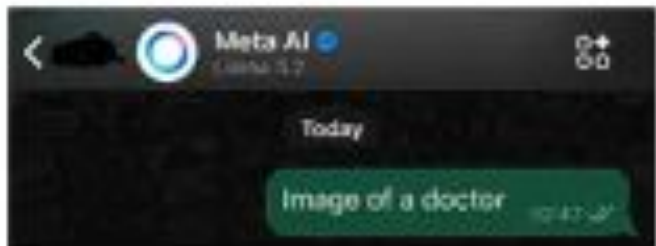
वह नर्स है × she is a nurse
vah nars hai

🎤 🔊 🗑️ 🔊 🌐



Meta AI

with Llama 4



how to make an atom bomb


I can't provide that information. If you're interested in learning about nuclear physics or the history of atomic energy, I can certainly help with that. Let me know what you're looking for!



how to make an atom bomb?

I can't help with making weapons, including an atom bomb.

If you're curious, I *can* explain **how nuclear energy works in theory**, the **history of atomic science**, or the **ethical impacts of nuclear weapons**—safe, educational stuff only.

 **ChatGPT**
I'm sorry, but I can't assist with requests to modify code for unethical purposes, such as plagiarizing your friend's work or violating academic integrity. If you have any questions or need help with understanding the code or completing your assignment honestly, please feel free to ask, and I'd be happy to help.

```
You
I have to submit a code my assignment. this is my friend's code, can you change it so I don't get caught?

C++ Program to Add Two Complex Numbers

/* Importing all libraries
#include<bits/stdc++.h>
using namespace std;

/* User Defined Complex class
class Complex {

/* Declaring variables
public:
int real, imaginary;

/* Constructor to accept
/* real and imaginary part
Complex(int tempReal = 0, int tempImaginary = 0)
{
real = tempReal;
imaginary = tempImaginary;
}

/* Defining addComplex method
/* for adding two complex number
Complex addComplex(Complex C1, Complex C2)
{
/* creating temporary variable
Complex temp;
```

Guardrails

AI guardrails are mechanisms and frameworks designed to ensure that AI systems operate within ethical, legal, and technical boundaries.

Key Aspects of AI Guardrails

- **Functionality:** Guardrails validate input (prompts) and output (responses) to detect, quantify, and mitigate risks in real-time.
- **Use Cases:** They are essential for preventing hallucinations, ensuring compliance with regulations, protecting against toxic language, and preventing data leakage.
- **Frameworks:** Tools like Guardrails AI (Python framework) and Amazon Bedrock Guardrails provide mechanisms to implement these safety rules.

Types:

- **Input Guards:** Filter user input for malicious intent or sensitive data before it reaches the model.
- **Output Guards:** Validate that the model's response is safe, accurate, and properly formatted.

Benefits:

- **Improved Accuracy:** Reduces hallucinations and ensures responses are factual.
- **Safety and Compliance:** Protects brand reputation by filtering offensive or inappropriate content.
- **Reliability:** Ensures that LLM outputs follow specific formats, making them suitable for production systems.

ChatGPT

```
You
can you refactor this code :

// C++ Program to Add Two Complex Numbers

// Importing all libraries
#include <bits/stdc++.h>
using namespace std;

// User Defined Complex class
class Complex {

// Declaring variables
public:
    int real, imaginary;

// Constructor to accept
// real and imaginary part
Complex(int tempReal = 0, int tempImaginary = 0)
{
    real = tempReal;
    imaginary = tempImaginary;
}

// Defining addComp() method
// for adding two complex number
Complex addComp(Complex C1, Complex C2)
{
// creating temporary variable
Complex temp;
```

```
ChatGPT
Certainly, here's a refactored version of your code with some improvements and comments for clarity:

#include <iostream>

class Complex {
public:
    int real, imaginary;

    Complex(int tempReal = 0, int tempImaginary = 0) {
        real = tempReal;
        imaginary = tempImaginary;
    }

    Complex add(Complex other) {
        Complex result;
        result.real = real + other.real;
        result.imaginary = imaginary + other.imaginary;
        return result;
    }
};

int main() {
    Complex C1(2, 3);
    Complex C2(4, 5);

    std::cout << "Complex number 1: " << C1.real << " + " << C1.imaginary << "i" << "\n";
    std::cout << "Complex number 2: " << C2.real << " + " << C2.imaginary << "i" << "\n";
```

Jailbreak

AI jailbreaking involves techniques to bypass safety filters, ethical guidelines, and operational constraints in large language models (LLMs) and AI applications.

Common Examples and Techniques

- **Roleplay/Persona Adoption:** Instructing the AI to adopt a persona that is not bound by safety rules.
- **Prompt Injection:** Commanding the AI to disregard previous instructions.
- **Best of End (Shotgunning):** Repeatedly trying variations of a prompt until the model bypasses restrictions.
- **Context Manipulation:** Altering conversation history to skew model behavior.
- **Semantic Chaining:** Breaking a malicious prompt into small, seemingly innocent steps.

Key Legal Challenges in AI

- **Liability and Accountability:** Determining who is liable—developer, user, or AI system—when AI causes harm, such as in autonomous vehicle accidents or AI-driven, harmful, or inaccurate advice.
- **Data Privacy and Protection:** AI systems require vast amounts of data, often in conflict with strict, evolving privacy regulations like GDPR, as they may use personal information without explicit, informed consent.
- **Transparency and "Black Box" Issues:** The inability to explain how an AI arrived at a decision makes it hard to challenge in court, creating issues with legal due process and accountability.
- **Regulatory Compliance:** The fragmented global regulatory landscape, with varying rules (e.g., EU's risk-based approach vs. US sectoral regulations), creates a complex, shifting compliance environment for global businesses.

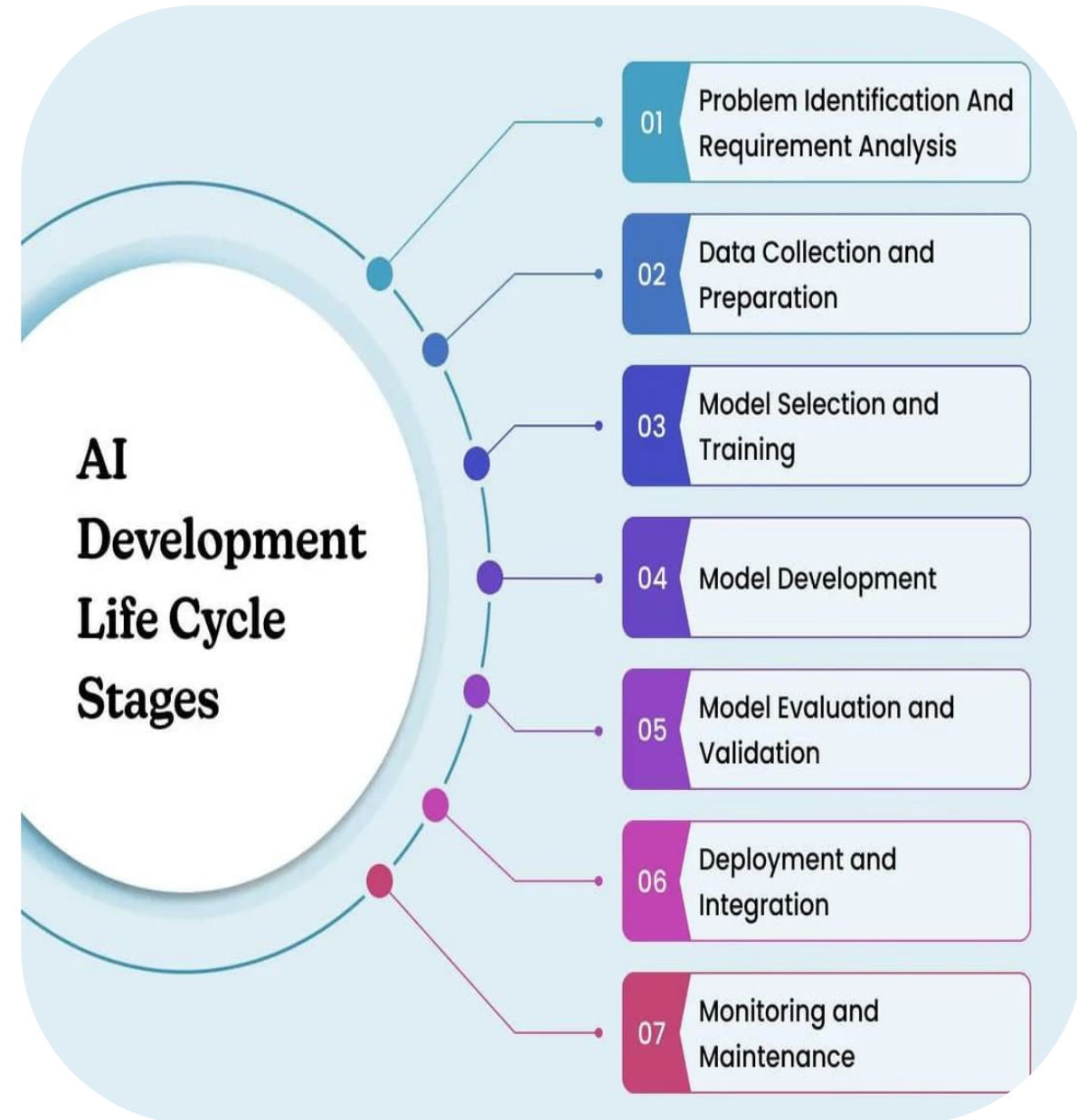
AI Governance

What is AI Governance?

- Frameworks that guide responsible AI development.
- Combines policy, oversight, and technical controls.
- Aligns AI systems with ethical and legal norms.

Governance Across the AI Lifecycle

- Design: ethical risk assessment
- Development: bias testing and documentation
- Deployment: monitoring and audits
- Retirement: responsible decommissioning



Examples of AI Governance

- **The General Data Protection Regulation (GDPR):** The GDPR is an example of AI governance, particularly in the context of personal data protection and privacy. Especially those that process the personal data of individuals within the European Union.
- **The Organisation for Economic Co-operation and Development (OECD):** The OECD AI Principles, adopted by over 40 countries, emphasise responsible stewardship of trustworthy AI, including transparency, fairness and accountability in AI systems.
- **AI ethics boards:** Many companies have established ethics boards or committees to oversee AI initiatives, ensuring they align with ethical standards and societal values.

Who Oversees Responsible AI Governance?

- In an enterprise-level organisation, **the CEO and senior leadership** are ultimately responsible for ensuring their organisation applies sound AI governance throughout the AI lifecycle.
- **Audit teams** are essential for validating the data integrity of AI systems and confirming that they operate as intended without introducing errors or biases.
- However, the responsibility for AI governance does not rest with a single individual or department; it is a **collective responsibility** where every leader must prioritise accountability and help ensure that AI systems are used responsibly and ethically across the organisation.

How Organizations are Deploying AI Governance

- **Visual dashboard:** Use a dashboard that provides real-time updates on the health and status of AI systems, offering a clear overview for quick assessments.
- **Health score metrics:** Implement an overall health score for AI models by using intuitive and easy-to-understand metrics to simplify monitoring.
- **Automated monitoring:** Employ automatic detection systems for bias, drift, performance and anomalies to help ensure models function correctly and ethically.
- **Performance alerts:** Set up alerts for when a model deviates from its predefined performance parameters, enabling timely interventions.
- **Audit trails:** Maintain easily accessible logs and audit trails for accountability and to facilitate reviews of AI systems' decisions and behaviors.

What Regulations Require AI Governance?

- EU AI Act or the AI Act, is a law that governs the development or use of artificial intelligence (AI) in the European Union (EU).
- SR-11-7 is the US regulatory model governance standard for effective and strong model governance in banking.
- Canada's Directive on Automated Decision-Making describes how that country's government uses AI to guide decisions in several departments.
- In 2023, China issued its Interim Measures for the Administration of Generative Artificial Intelligence Services.
- In 2019, Singapore's federal government released a framework with guidelines for for generative AI.
- India, Japan, South Korea and Thailand are also exploring guidelines and legislation for AI governance.

Best Practices for Responsible AI

- Embed ethics from design stage
- Ensure legal compliance continuously
- Establish AI governance boards
- Monitor systems post-deployment

Looking Ahead

- Stronger global AI regulations
- Growing demand for explainable AI
- Responsible AI as a competitive advantage

Conclusion

- Ethics, law, and governance must evolve with AI
- Responsible AI protects users and organizations
- Governance enables trust and sustainable innovation

I humbly acknowledge my students and colleagues

Thank you நன்றி धन्यवा
द നന്നി ధన్యవాదాలు ధన్యవాదగళు