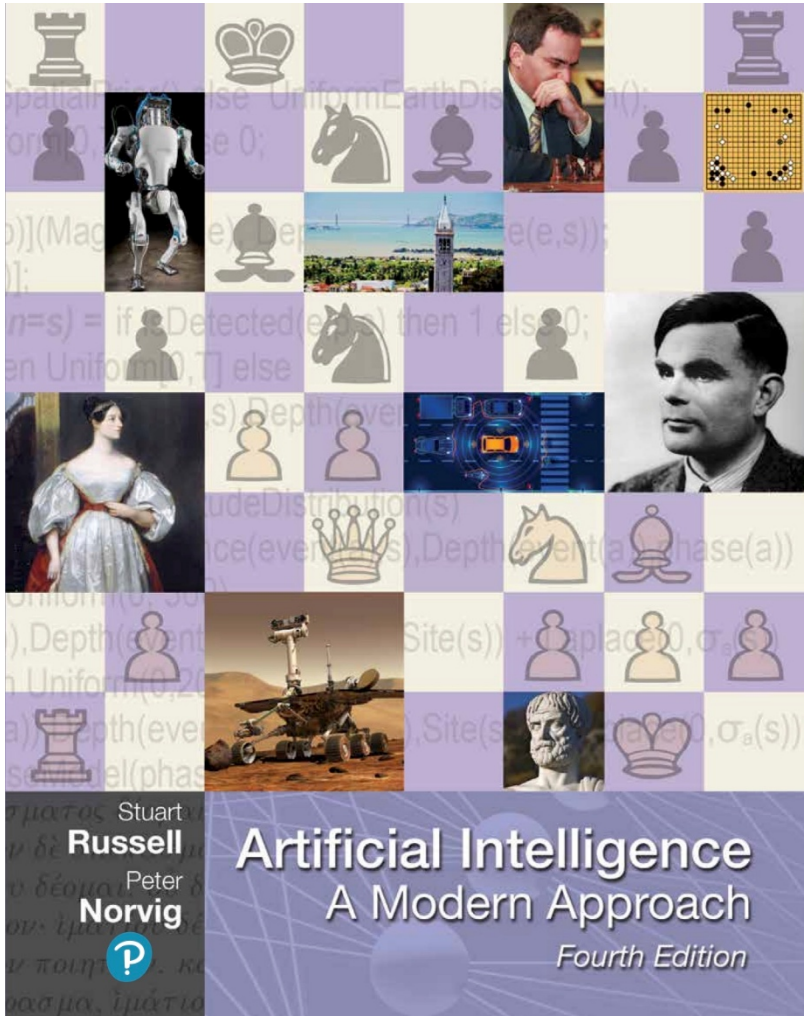
AI Foundations for Engineers

Unit – 1 (07 Hours)

Introduction: What is AI? Acting humanly: The Turing test approach,
Thinking humanly: The cognitive modelling approach,
Thinking rationally: The "laws of thought" approach,
Acting rationally: The rational agent approach;
The foundations of AI: Mathematics, Economics, Neuroscience, Psychology, Computer Engineering; The State of the Art; Risks and Benefits of AI

General Introduction to Responsible AI: What is Responsible AI? Why it is important

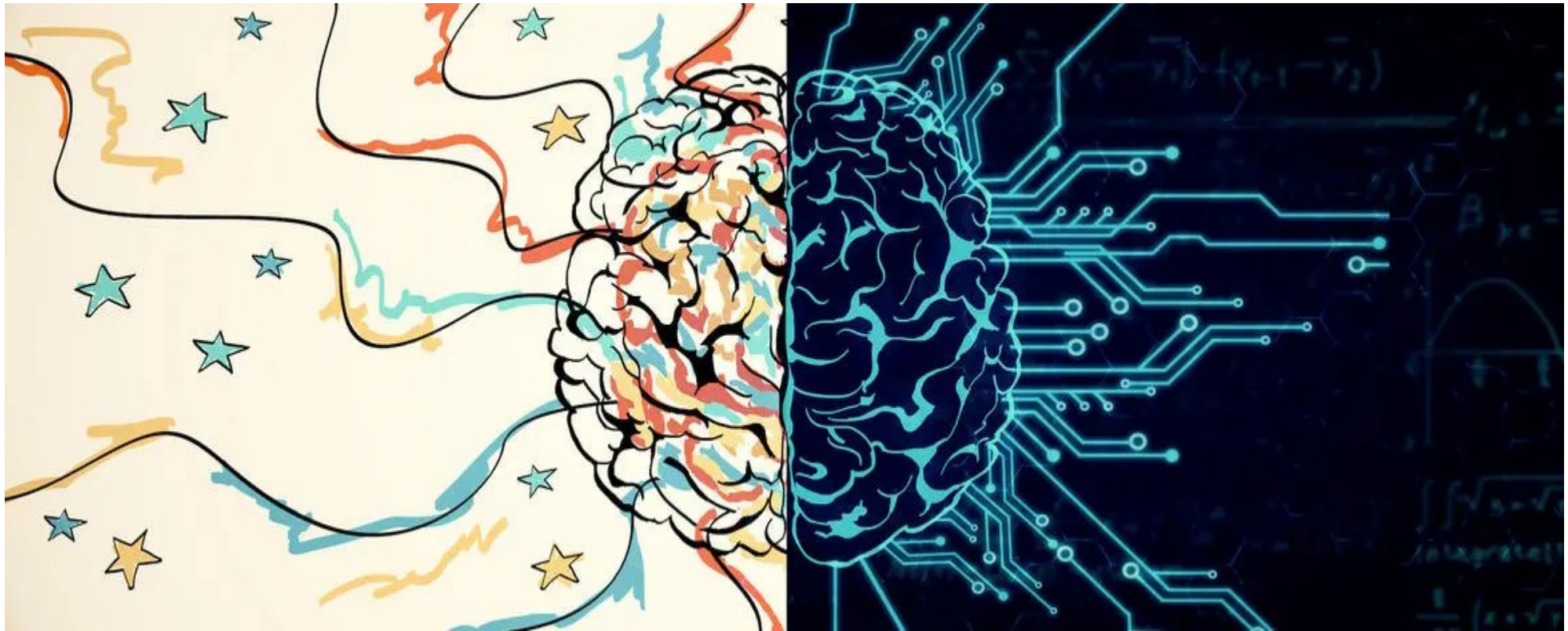
Intelligent Agents: Agents and Environments, The concept of Rationality, The Nature of Environments



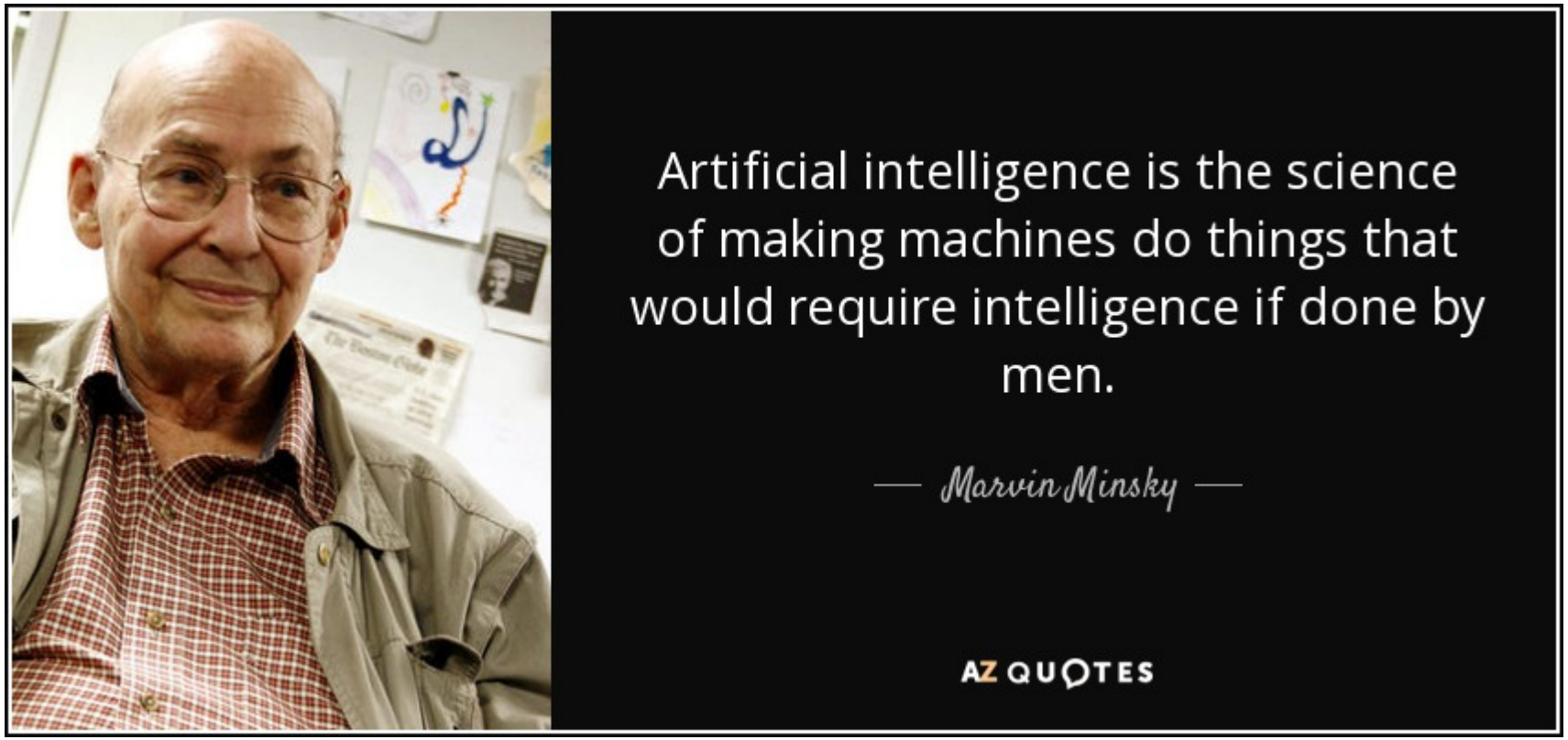
Artificial Intelligence?

Artificial Intelligence (AI) is a comprehensive area that comprises creating machines or software which exhibit *intelligent* behavior. In simple terms, AI systems are built to execute the tasks that needs human ingenuity, such as understanding language, learning from experience, making decisions, or solving problems.

AI is “the study of **agents** that receive *percepts* from the environment and perform *actions*” by Russell and Norvig



Artificial Intelligence

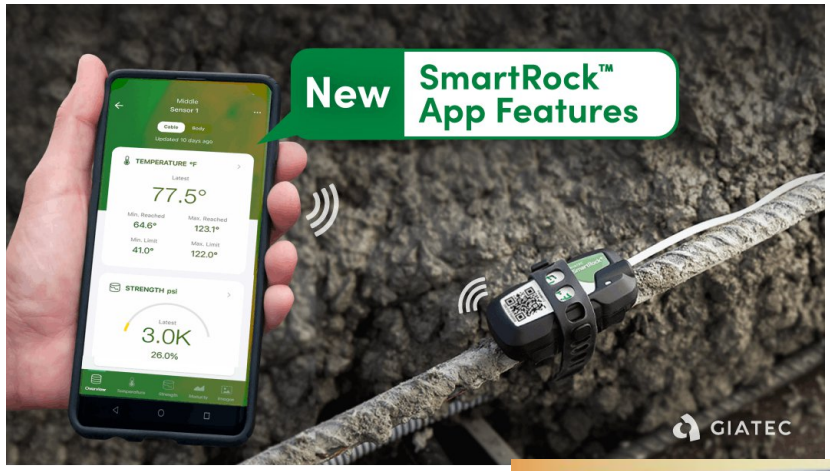


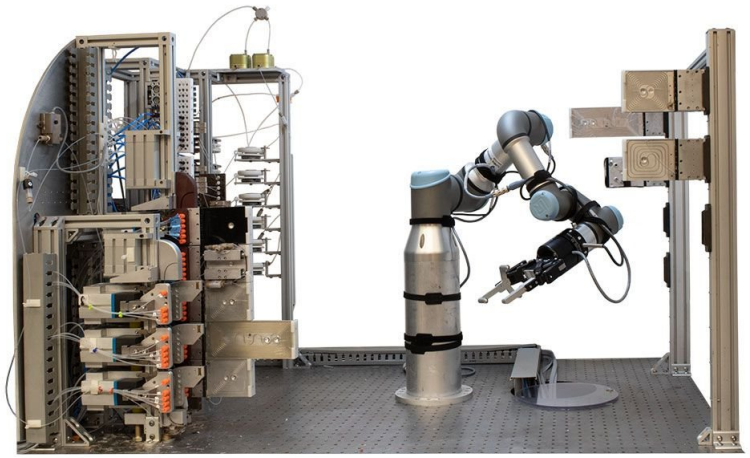
Marvin Minsky (1927-2016) was a pioneering cognitive scientist and co-founder of the MIT Media Lab, widely regarded as one of the founding fathers of artificial intelligence.

AI Focus

- AI focuses on building intelligent machines that can effectively and safely navigate **novel situations.**

- One of the core aspects of AI is its ability to **make decisions and solve problems.**





AI Focus

Can you think of an innovative application in your discipline that could benefit from AI intervention?

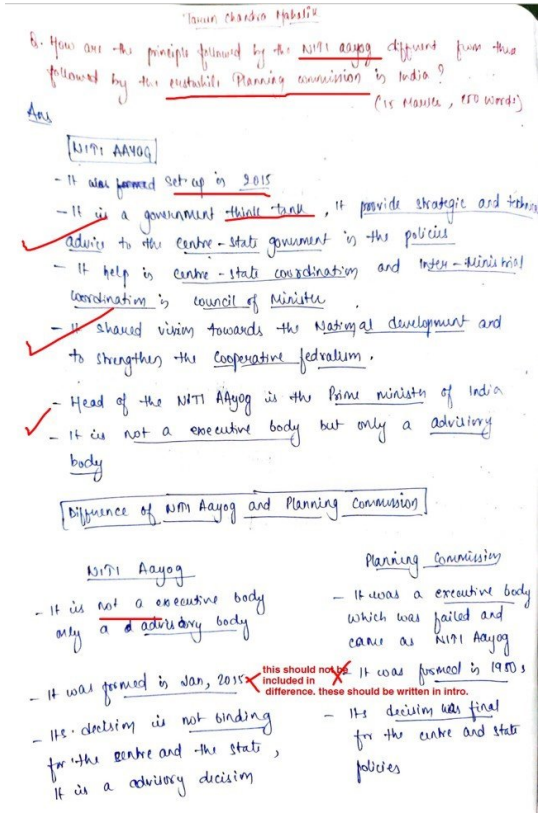
Reflection Note of 100 words

Guess the word…?

The capability to make decisions and solve problems hinges on the concept of

_ _t_ _n_ _ _ _y

Rationality in our profession-examples



- It was formed set up in 2015
- It is a government think tank, It provide strategic and tech... advice to the centre - state government in the policies
- It help in centre - state coordination and inter - ministrial coordination in council of Minister.
- It shared vision towards the National development and to strengthen the cooperative federalism.
- Head of the NITI AAyog is the Prime minister of India
- It is not a executive body but only a advisory body

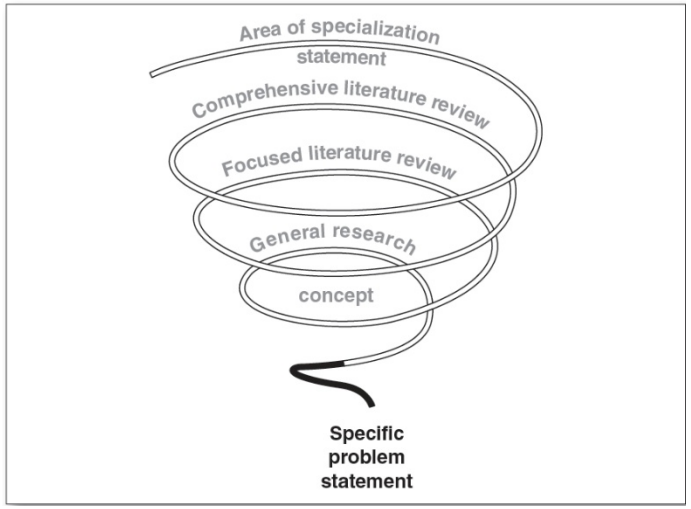
Difference of NITI Aayog and Planning Commission

NITI Aayog	Planning Commission
- It is not a executive body only a advisory body	- It was a executive body which was failed and came as NITI Aayog
- It was formed in Jan, 2015	- It was formed in 1950's
- The decision is not binding for the centre and the state, It is a advisory decision	- Its decision was final for the centre and state policies

this should not be included in difference. these should be written in intro.

Grading and Evaluation

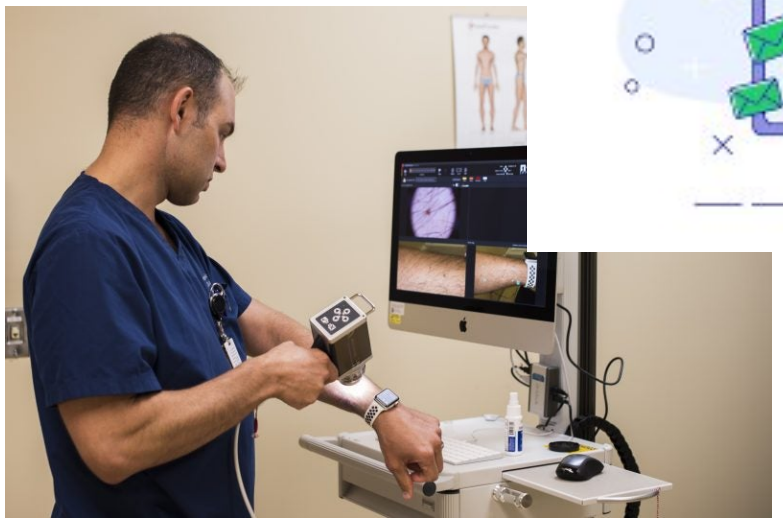
Choosing Teaching Methods for a Course



Finalizing Research Problem

Rationality

- It is the formal definition of intelligence
- Intelligence is nothing but doing right thing always
- It is internal thought processes and reasoning
- Intelligence is practical, ethical, and consistent in action, not just theoretical knowledge



Rationality

- It is internal thought processes and reasoning

Rationality is the ability of an artificial agent to make decisions that maximize its performance based on the information it has and the goals it seeks to achieve.

In essence, a rational AI system aims to choose the best possible action from a set of alternatives to achieve a specific objective.

This involves logical reasoning, learning from experiences, and adapting to new situations.

The robotaxi–ambulance example illustrates how AI's "rationality" (rule-following optimization) can directly **conflict with human rationality (value-sensitive, context-aware prioritization of lives over rules)**.

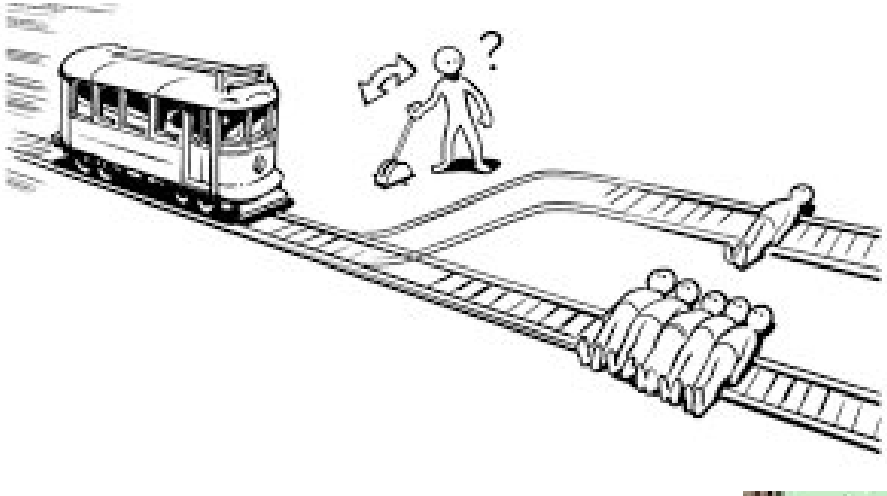
It highlights the need to design AI systems that incorporate **ethical reasoning, context-awareness, and value alignment**—not just rigid optimization.

What Happened in the Example

- A robotaxi stops in the middle of the road because its sensors and ruleset tell it **not to break traffic laws** or move unpredictably.
- An ambulance behind it is delayed, compromising an **emergency response**.
- From the robotaxi's programmed perspective, it acted "rationally" (follow the rules, avoid risk).
- From society's perspective, the outcome was **irrational and harmful** (delayed life-saving intervention).
- This reveals why **encoding values into AI rationality is critical**.

Rationality

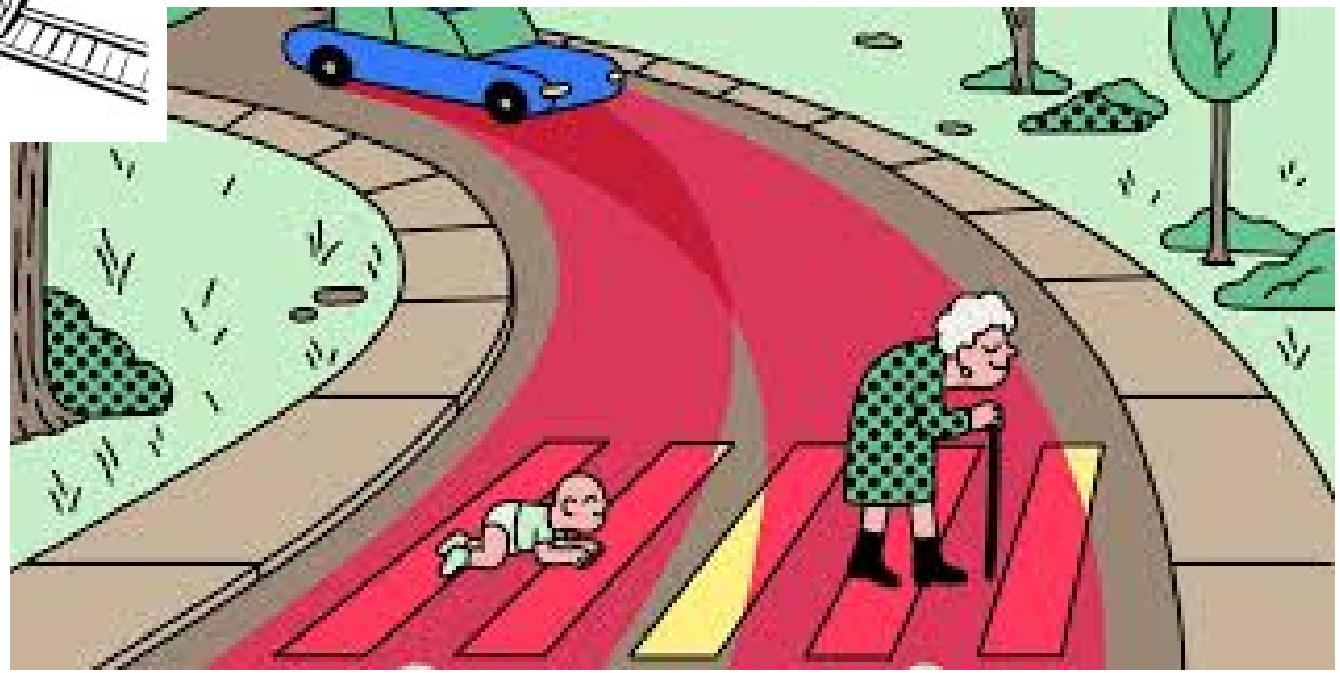
- **Practical intelligence** solves real-world problems.
- **Ethical intelligence** promotes responsible use of abilities.
- **Consistent intelligence** means doing the right thing repeatedly.



Trolley Problem

a classic thought experiment in ethics and psychology that presents a moral dilemma

Emotional Reasoning
v/s
Rational Reasoning



Should a self-driving car kill the baby or the grandma?


MakeAGIF.com

Emotional Reasoning
v/s
Rational Reasoning

Lifeboat Dilemma

A lifeboat dilemma is a **moral decision-making problem**:

•Imagine a lifeboat with limited capacity.
•More people need rescue than the boat can carry.

•Someone must decide **who gets saved** and **who gets left behind**.

This scenario forces us to weigh **competing values** (e.g., save the young vs. the old, strongest vs. weakest, more lives vs. fewer).

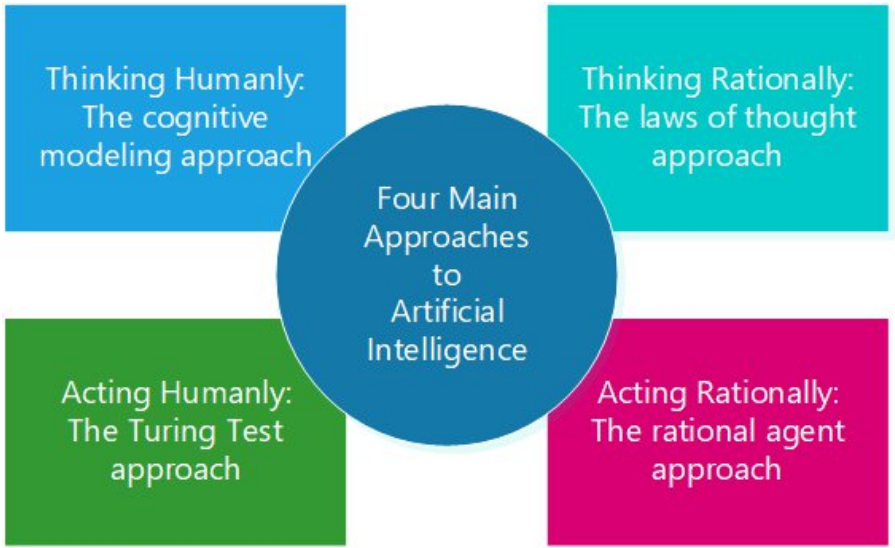
The lifeboat dilemma is a metaphor for AI rationality because it shows how **limited resources, conflicting objectives, and moral trade-offs** force difficult but necessary rational decisions.

What Is AI?

- Historically, researchers have pursued several different versions of AI.
- Some have **defined intelligence** in terms of fidelity to *human* performance, while others prefer an abstract, formal definition of intelligence called **rationality**—loosely speaking, doing the "right thing."

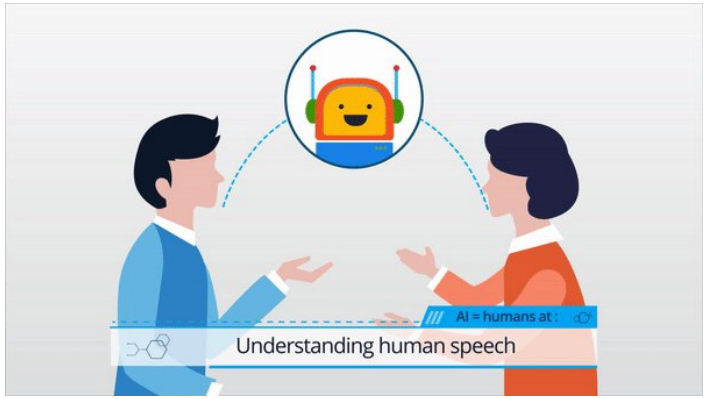
- The subject matter itself also varies- some consider intelligence to be a property of internal *thought processes* and *reasoning*, while others focus on intelligent *behavior*, an external characterization.
- From these two dimensions—**Human vs. Rational** and **Thought vs. Behavior** hence, there are four possible combinations.

1. **Acting humanly: The Turing test approach**
2. **Thinking humanly: The cognitive modeling approach**
3. **Thinking rationally: The "laws of thought" approach**
4. **Acting rationally: The rational agent approach**



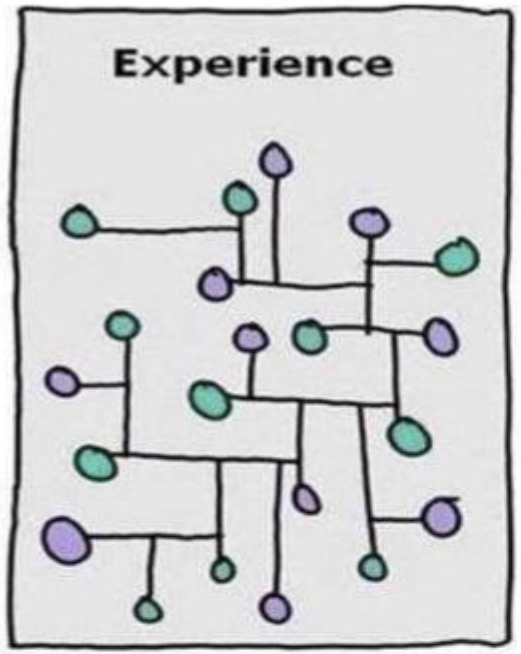
Acting humanly: The Turing test approach

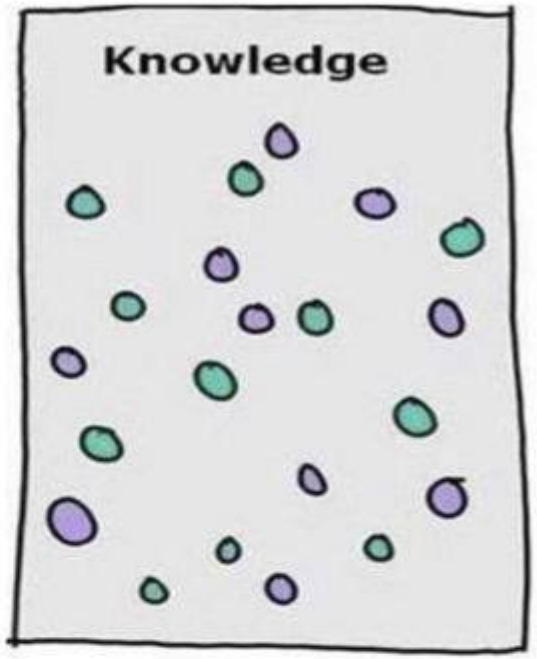
- The Turing test, proposed by Alan Turing (1950), was a thought experiment => **"Can a machine think?"**
- The computer would need the following capabilities:
  - Natural language processing for human communication
  - Knowledge representation for information storage
  - Automated reasoning for answering questions and drawing conclusions
  - Machine learning for adaptation and pattern recognition



Experience refines knowledge through practice and feedback.

AI gains “experience” via machine learning — training on data, improving performance with exposure, and adapting from feedback



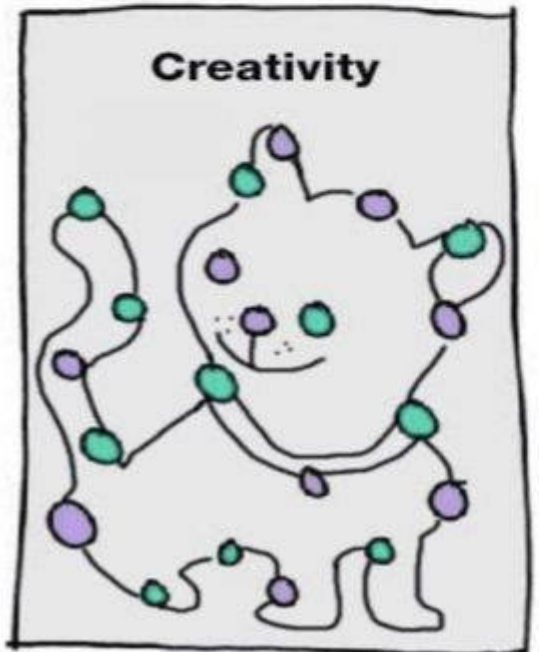


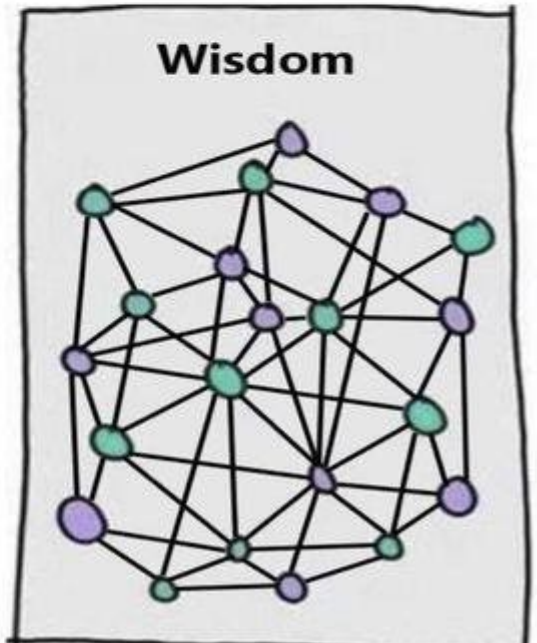
Knowledge is what you know and understand, gained through learning

In Artificial Intelligence (AI), knowledge refers to the information that an intelligent system stores, uses, and reasons with to make decisions or perform tasks.

Creativity involves producing novel, valuable ideas or artifacts by combining knowledge and experience.

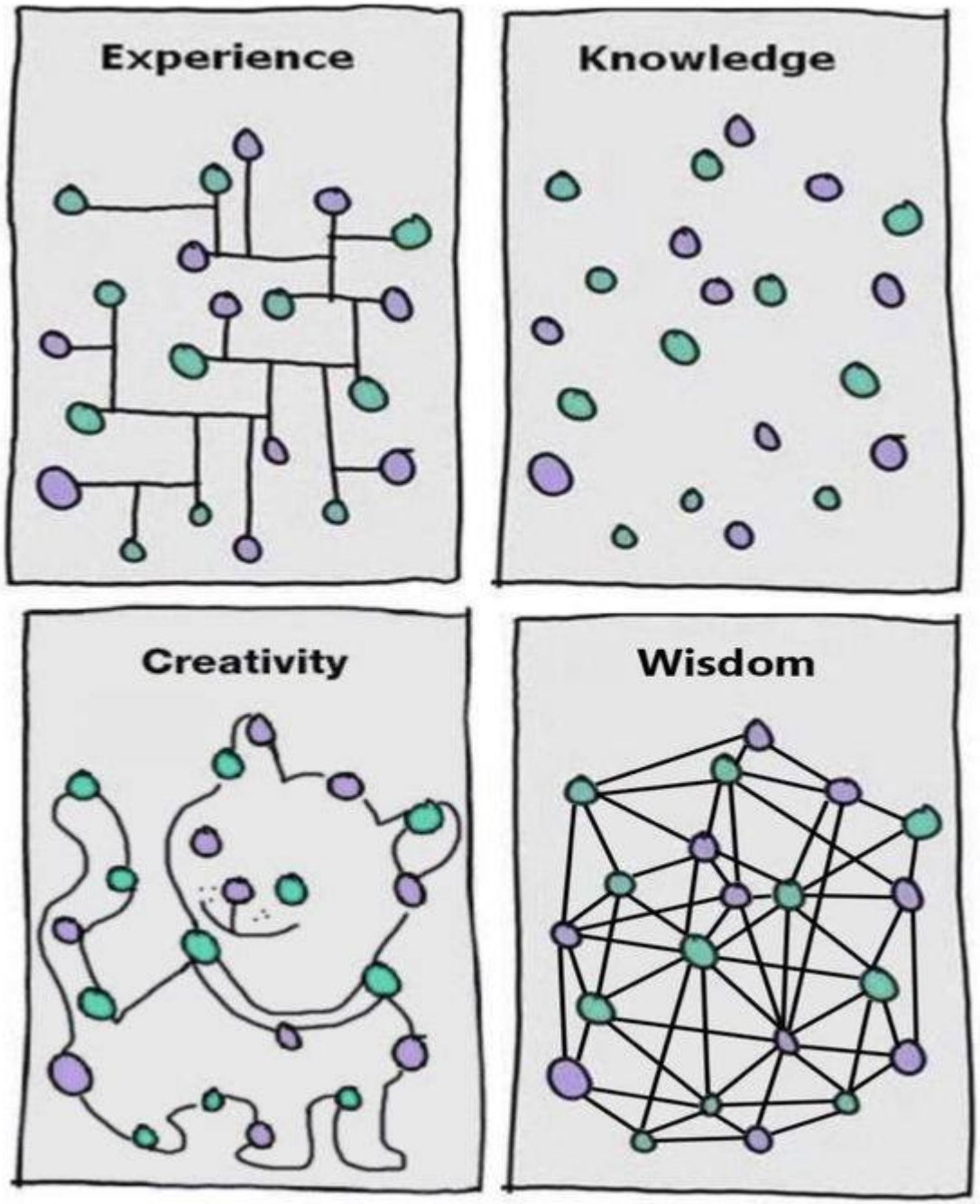
Generative AI demonstrates creativity by synthesizing new images, music, or text. AI creativity is pattern-driven rather than intuitive

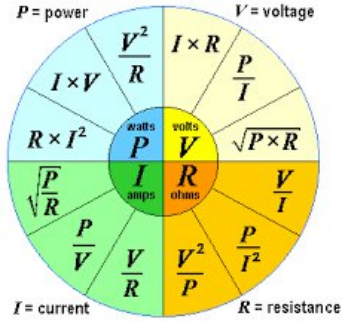




Wisdom is the ability to apply knowledge and experience ethically, contextually, and for long-term good.

True wisdom is the hardest for AI. Current AI lacks intrinsic values or ethics — it relies on human-aligned frameworks, ethical AI principles, and governance mechanisms to ensure decisions serve societal good.


Experience
Knowledge
Creativity
Wisdom


P = power
V = voltage
I = current
R = resistance
watts
volts
amps
ohms

Knowledge Representation

Knowledge Representation encodes information for computers to understand, reason, and act on.

AI does not use *random* knowledge that is unstructured or meaningless.

Knowledge Representation

- Logical Representation : $\forall x\ Cat(x) \rightarrow Mammal(x)$

Try This:

Everybody loves somebody

Knowledge Representation

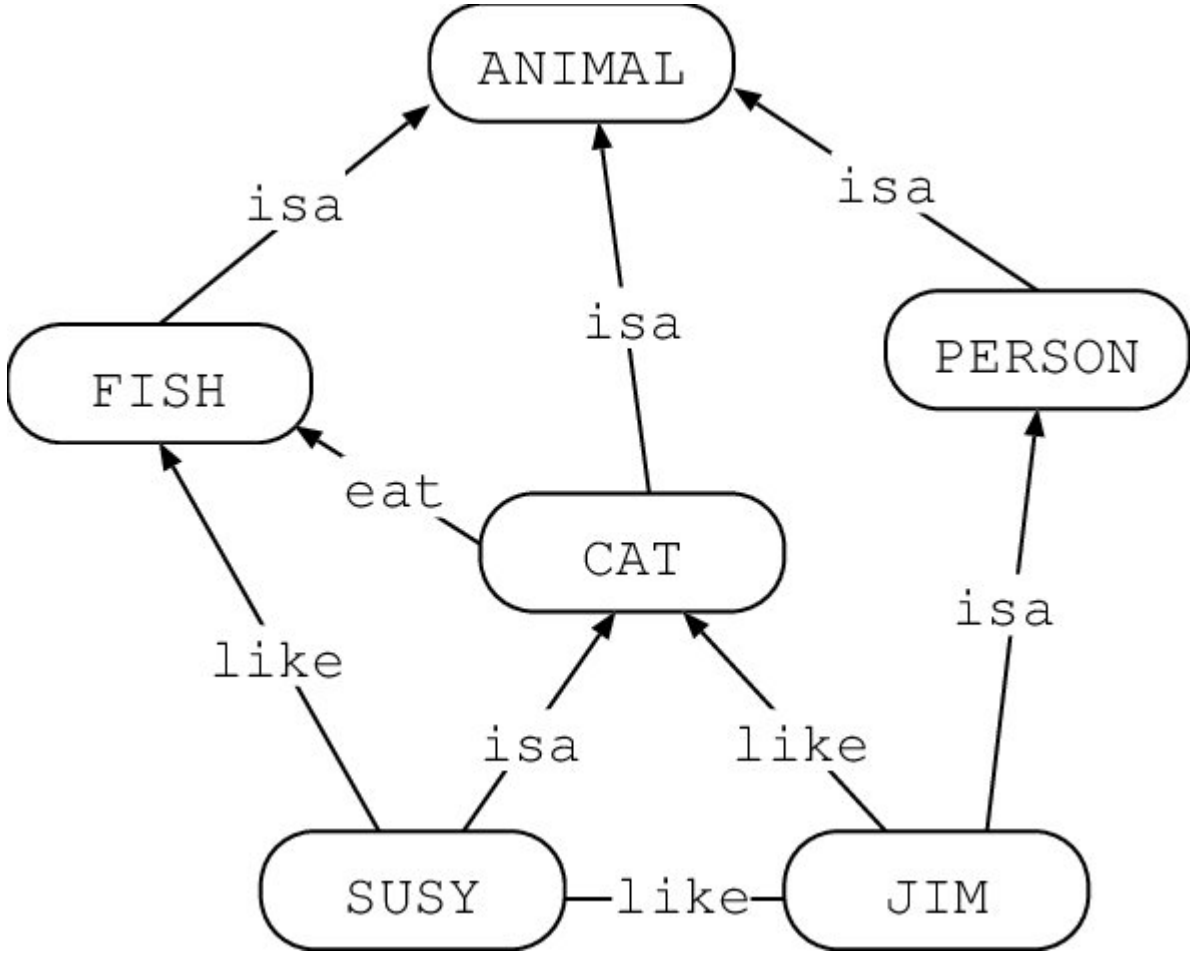
- ∀x∃y Loves(x,y)

Try This:

It is not cold but it is cloudy, where p: it is cold, and q: it is cloudy.

Knowledge Representation

- Semantic Networks (Graphs)



Try This:

Earth revolves around the Sun

Knowledge Representation

- Frames (Structured Data)

Frame: Car

Slots (Attributes)	Values
Type	Sedan
Brand	Toyota
Model	Corolla
Color	Red
Year	2022
Engine Type	Petrol
Number of Doors	4
Owner	Mr. Rangesh

Try This:

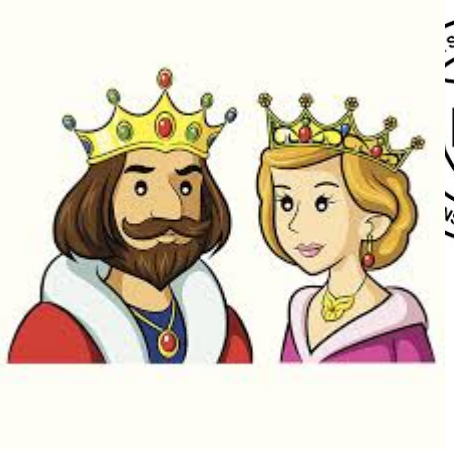
Create a 10 attributes Data Frame for RVCE

Knowledge Representation

- Production Rules (If-Then Rules):
  - **IF** animal has fur **AND** animal meows **THEN** animal is likely a cat.

- IF traffic density > 60 vehicles/min AND average speed < 15 km/h AND signal cycle time > 90s THEN optimize signal timing.

Knowledge Representation



- Neural Representations (Vector Embeddings): AI stores knowledge implicitly as vectors in high-dimensional space.

king → [0.25, 0.85, 0.60]

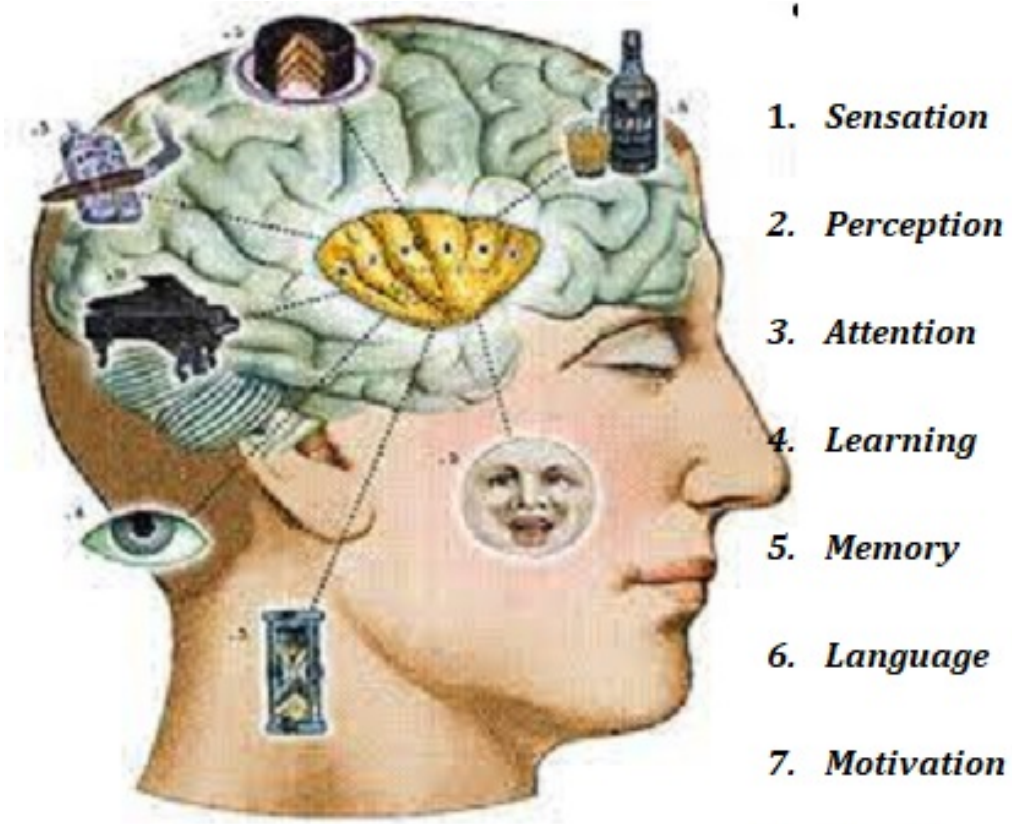
queen → [0.30, 0.80, 0.65]

man → [0.20, 0.75, 0.50]

woman → [0.25, 0.70, 0.55]

```
array([-0.5968882 , -0.33086956, -0.32643065, -0.3670732 ,  0.628059  ,
       -0.3692328 , -0.37902787, -0.12308089, -0.38124698, -0.03940517,
        0.2260839 ,  0.10852845, -0.2873811 , -0.42781743,  0.06604357,
       -0.07114276, -0.29775023, -0.99628943, -0.54497653, -0.11718027,
       -0.15935768,  0.09587188, -0.2503798 ,  0.06768776,  0.3311586 ,
        0.43098116,  0.06936899,  0.24311952,  0.14515282,  0.19245838,
        0.10462623, -0.45676082,  0.5662387 ,  0.69908774,  0.48064467,
        0.27378514, -0.45430255,  0.17282294, -0.40275463, -0.38083532,
        0.47487524,  0.31950948, -0.1109335 ,  0.2165357 ,  0.034114  ,
        0.05689918,  0.20939653,  0.15209009, -0.24204595,  0.03478364,
        0.1616051 , -0.5827333 , -0.47017908,  0.26226178, -0.11884775,
        0.40180743, -0.5173988 , -0.19270805,  0.660391  , -0.24518126,
       -0.42860952, -0.22274768,  0.4887834 ,  0.49302152,  0.38799986,
       -0.041193  , -0.38600504, -0.37632987,  0.04570564,  0.50462466,
       -0.14396502,  0.33490512, -0.15964787, -0.21363072, -0.25445372,
        0.52389127,  0.5747422 , -0.25075617, -0.5339069 ,  0.2582965 ,
       -0.16139959,  0.09748188,  0.04540966, -0.27768216, -0.51260656,
       -0.06189002, -0.54032195, -0.21863565,  0.06233869,  0.13287479,
        0.49741864,  0.1772418 ,  0.02064824, -0.04775626, -0.16804916,
        0.4643644 ,  0.5546319 ,  0.68051434,  0.7790246 ,  0.5617202 ],
      dtype=float32)
```

Thinking humanly: The cognitive modeling approach



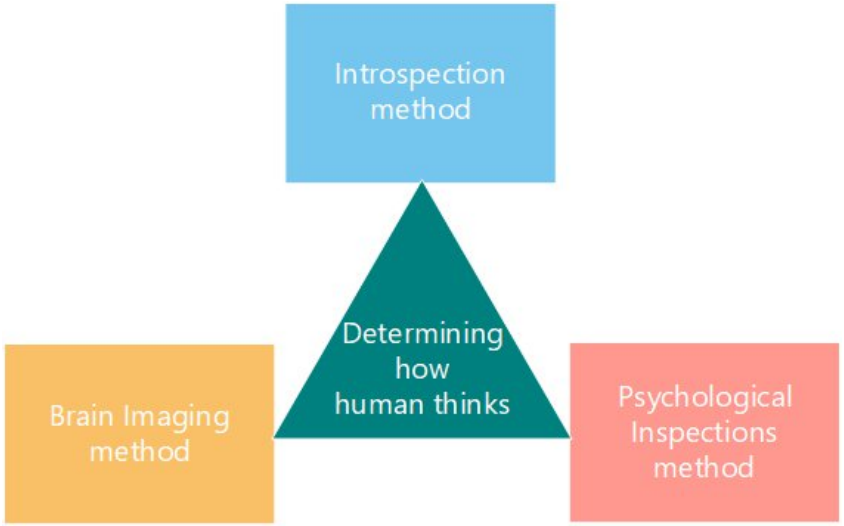
1. *Sensation*
2. *Perception*
3. *Attention*
4. *Learning*
5. *Memory*
6. *Language*
7. *Motivation*
8. *Emotion*

Mental Processes in Psychology

Thinking humanly: The cognitive modeling approach

We can learn about human thought in three ways:

- **introspection**—trying to catch our own thoughts as they go
- **psychological experiments**—observing a person in action;
- **brain imaging**—observing the brain in action.



The interdisciplinary field of **cognitive science** brings together computer models from AI and experimental techniques from psychology to construct precise and testable theories of the human mind.

Once we have a sufficiently precise theory of the mind, it becomes possible to express the theory as a computer program. If the program's input–output behavior matches correspond ing human behavior, that is evidence that some of the program's mechanisms could also be operating in humans.

Thinking humanly: The cognitive modeling approach

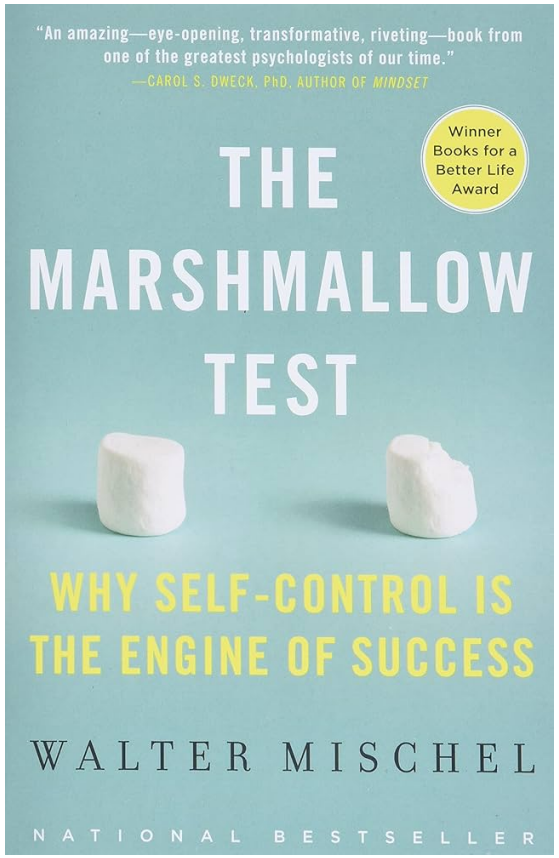
- **Introspection**—trying to catch our thoughts as they go by;

- **Feeling Anxious**:

"My heart's racing… I'm thinking the worst… Is the situation really dangerous, or is my mind exaggerating?"

Thinking humanly: The cognitive modeling approach

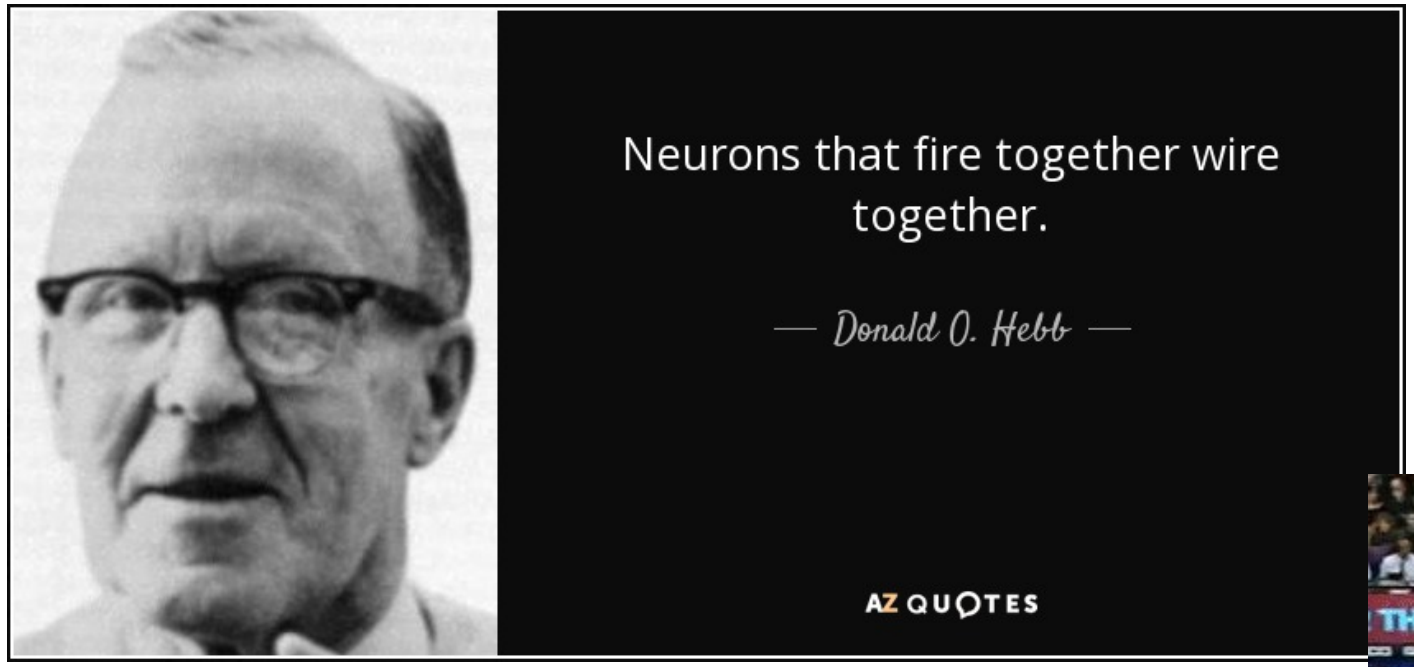
- **Psychological experiments**—observing a person in action;



Thinking humanly: The cognitive modeling approach

- **Brain imaging** —observing the brain in action.

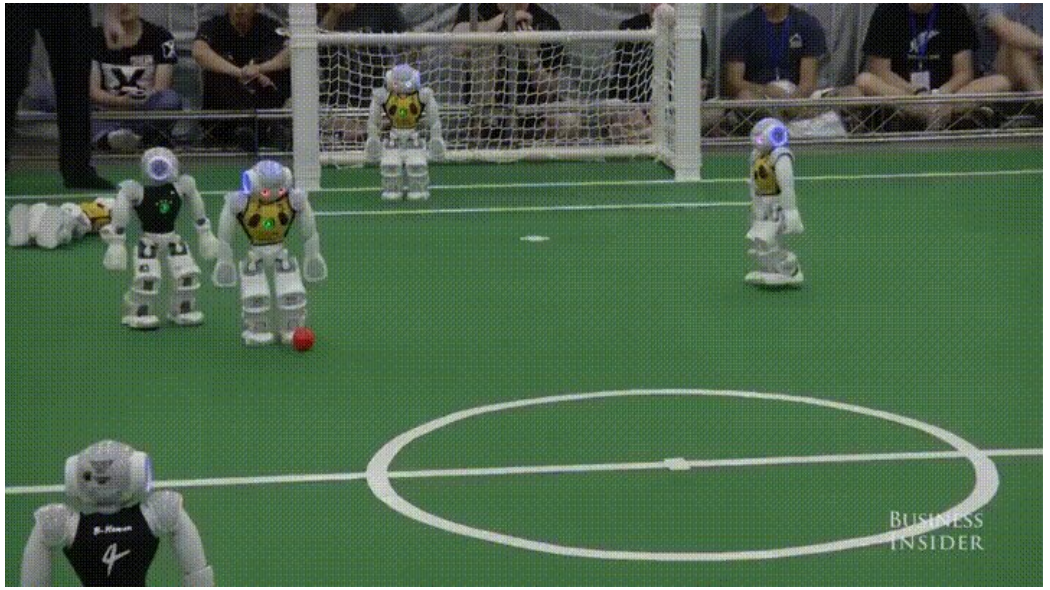
Thinking humanly: The cognitive modeling approach



Thinking humanly: The cognitive modeling approach

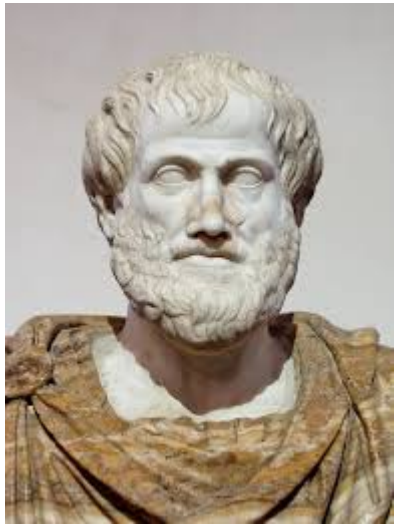
- **Cognitive science** combines AI computer models and psychological experiments to develop testable theories about the human mind.

Idea: Build AI that mimics how humans actually think—step by step—using findings from cognitive psychology and neuroscience. We don't just want the right answer; we want the **same process and mistakes** humans make (reaction times, slips, learning curves).



Thinking rationally: The “laws of thought” approach

The *“laws of thought”* approach in Artificial Intelligence (AI) comes from classical logic and philosophy. It is based on the idea that **intelligent reasoning means following strict logical rules**—just as humans do when they think rationally. These rules are usually expressed in terms of *formal logic* (like Aristotle’s syllogisms or modern propositional and predicate logic).



- Rational thinking = following well-defined **laws of logic**.
- If the premises (inputs) are true, the reasoning process **must lead to true conclusions**.

This approach underpins early AI systems such as ***expert systems*** **and** ***rule-based systems***.

Syllogisms provided patterns for argument structures that always yielded correct conclusions when given correct premises.

Thinking rationally: The “laws of thought” approach

A **syllogism** is a kind of logical argument that applies deductive reasoning to arrive at a conclusion based on two propositions that are asserted or assumed to be true.

It follows a fixed structure:

1.Major Premise – a general statement.

2.Minor Premise – a specific statement related to the general one.

3.Conclusion – logically follows from the two premises.

- **Pattern**:
  All A are B.
  All B are C.
  Therefore, all A are C.
- **Example**:
  All humans are mammals.
  All mammals are animals.
  Therefore, all humans are animals.

Try this:
No A are B.
All C are A.
Therefore, no C are B.

Thinking rationally: The “laws of thought” approach

Types of Syllogisms

1.Categorical syllogism – deals with categories.

1. Example:
    1. All birds have wings.
    2. A sparrow is a bird.
    3. So, a sparrow has wings.

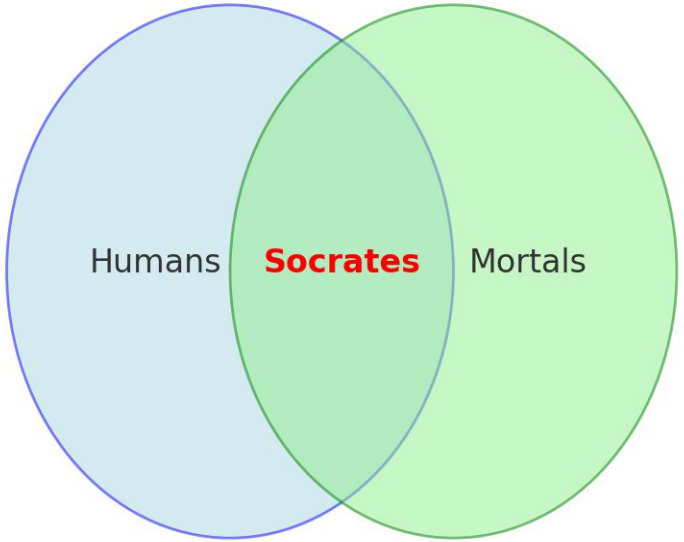
2.Hypothetical syllogism – uses “if…then” statements.

1. Example:
    1. If it rains, the ground will be wet.
    2. If the ground is wet, the match will be canceled.
    3. Therefore, if it rains, the match will be canceled.

3.Disjunctive syllogism – uses “either…or” statements.

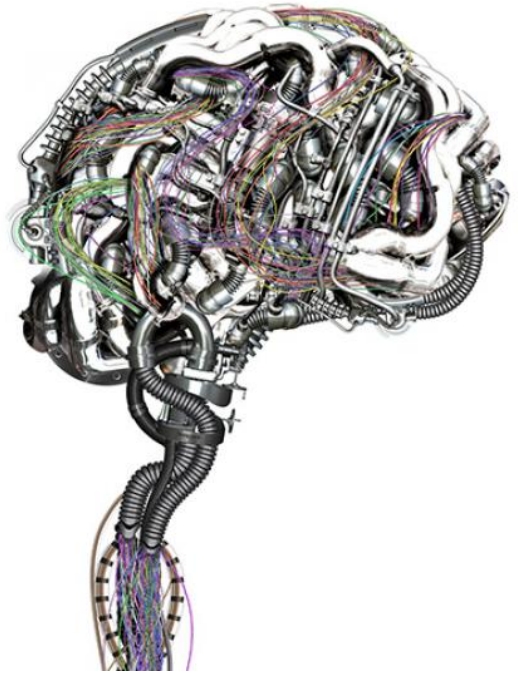
1. Example:
    1. Either the light is on or the room is dark.
    2. The light is not on.
    3. Therefore, the room is dark.

Syllogism Example: All humans are mortal
Socrates is a human → Socrates is mortal



Thinking rationally: The “laws of thought” approach

- Logic: the study of the mind's operations
- By 1965, programs could solve problems in logical notation.
- Logicist tradition in AI built intelligent systems.
- Logic requires world knowledge (conditioned).
- Probability theory enables reasoning with uncertainty.



Limitations

•**Rigidity**: Only works when knowledge can be expressed in logical rules.

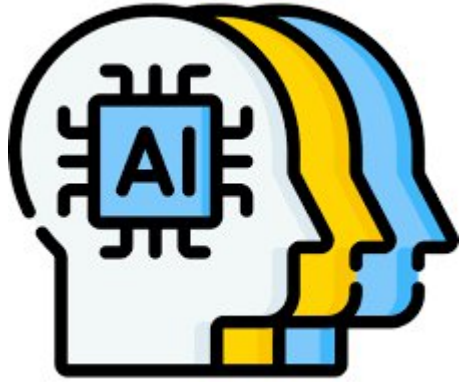
•**Incomplete knowledge**: If facts are missing, reasoning fails.

•**Uncertainty**: Real-world situations often require probability (handled better by modern AI like machine learning).

The “laws of thought” approach means programming machines to follow formal logical rules, ensuring rational conclusions—just like classical reasoning in mathematics and philosophy.

Acting rationally: The rational agent approach

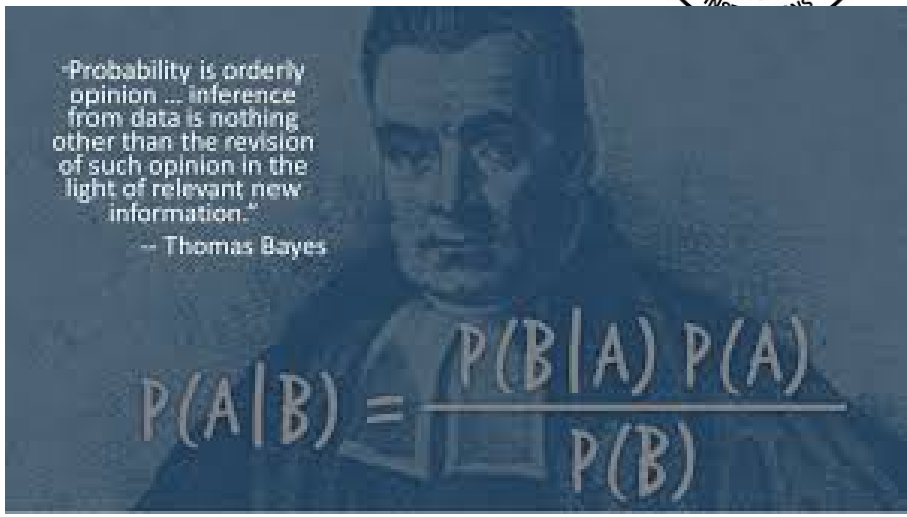
- **An agent acts.**
- Computer agents operate autonomously, perceive their environment, persist over time, adapt, and pursue goals.
- A rational agent seeks the best outcome, or the best expected outcome under uncertainty.
- The “laws of thought” in AI focus on correct inferences.
- To act rationally, deduce the best action, and act on it.
- **Skills for the Turing test enable rational action.**



Foundations of AI (Philosophy)

- **Philosophy**
  - Can formal rules be used to draw valid conclusions?
  - How does the mind arise from a physical brain?
  - Where does knowledge come from?
  - How does knowledge lead to action?
- **Dualism**: Belief in two distinct elements, like mind and body.
- **Materialism**: Everything, including the mind, is made of physical matter and energy.
- **Empiricism**: Knowledge comes from sensory experience.
- **Induction**: General rules emerge from repeated associations.
- **Logical Positivism**: Knowledge can be characterized by logical theories.
- **Confirmation Theory:** Analyzes knowledge acquisition by quantifying belief.

Foundations of AI (Mathematics)



- **Mathematics**
  - What are the formal rules to draw valid conclusions?
  - What can be computed?
  - How do we reason with uncertain information?
- **Formal Logic**
  - Propositional logic/Boolean logic/First-order logic
- **Probability**
  - Probability theory generalizes logic to uncertain situations, which is very important for AI.
  - Thomas Bayes (1702–1761) proposed a rule for updating probabilities with new evidence, which is essential for AI systems.
- **Statistics**
  - The formalization of probability, combined with the availability of data, led to Statistics
  - Ronald Fisher is regarded as the first modern statistician, blending probability, experimental design, data analysis, and computing.

Foundations of AI (Economics)

- **Economics**
  - How should we make decisions in accordance with our preferences?
  - How should we do this when others may not go along?
  - How should we do this when the payoff may be far in the future?
- **Decision Theory**
  - Integrates probability and utility theories to create a structured framework for making individual decisions under uncertainty, where probabilistic descriptions accurately reflect the decision-maker's environment.
- **Game Theory**
  - It is the study of mathematical models of strategic interaction among rational decision-makers (the actions of one player can significantly affect the utility of another).
- **Multi-agent Systems**
  - In AI, decisions involving multiple agents are studied under the heading of multi-agent systems
- **Operations Research**
  - How to make rational decisions when payoffs from actions are delayed and depend on a series of sequential actions.
- **Markov-Decision Process**
  - It is a framework for modeling decision-making in situations where outcomes depend on both current actions and probabilistic state transitions.

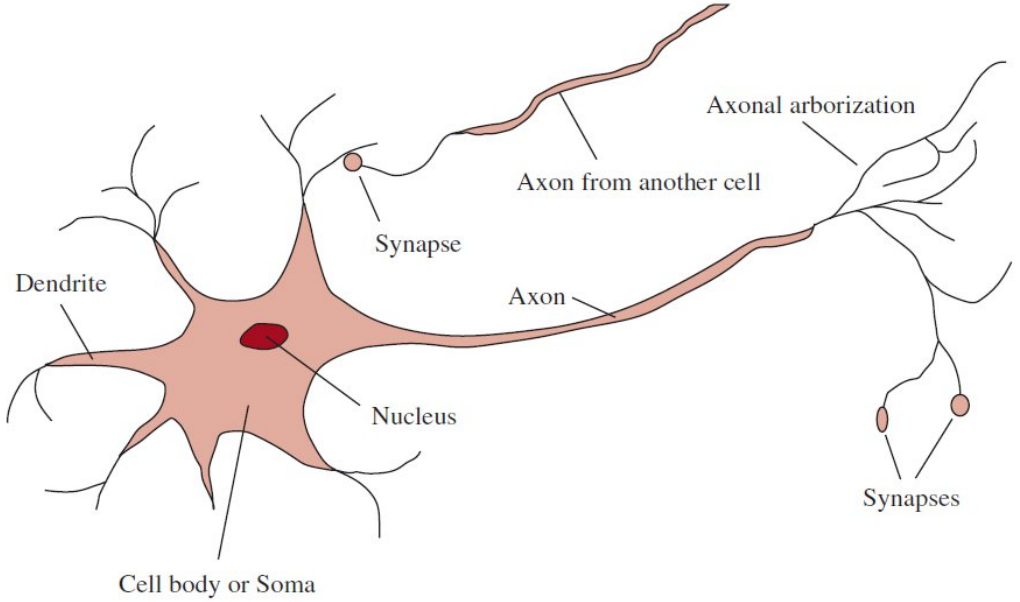
Guess what is used here?

Guess what is used here?

Guess what is used here?

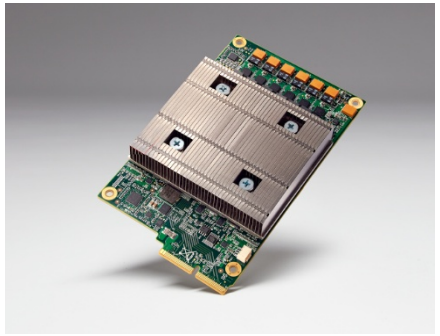
Foundations of AI (Neuroscience)

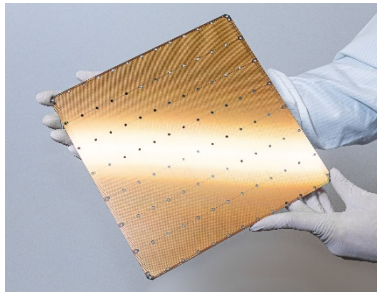
- **Neuroscience**
  - It is the study of the brain's nervous system
  - The brain consists of nerve cells called Neurons (a collection of simple cells can lead to thought, action, and consciousness)
  - Cognitive psychology views the brain as an information-processing device.
  - Intelligence augmentation (IA not AI) - computers should augment human abilities rather than automate away human tasks

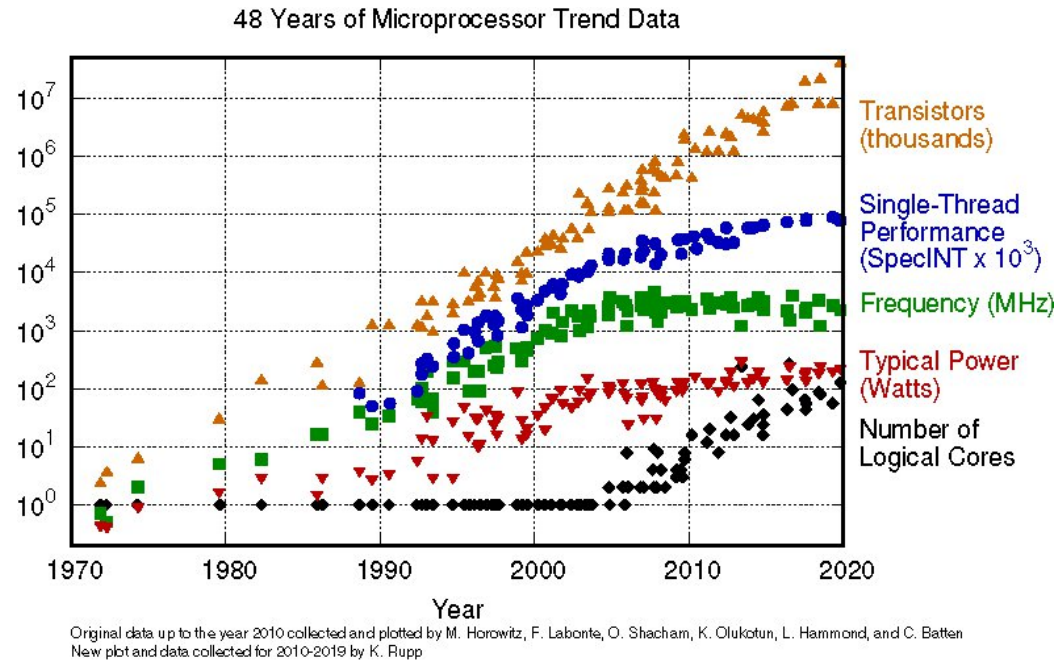


Foundations of AI (Computer Engineering)

- **Computer Engineering**
  - Increase in speed and capacity, and a decrease in price (Moore's law)
  - Hardware tuned for AI applications, such as the graphics processing unit (GPU), tensor processing unit (TPU), and wafer-scale engine (WSE)
  - Quantum computing holds out the promise of far greater accelerations







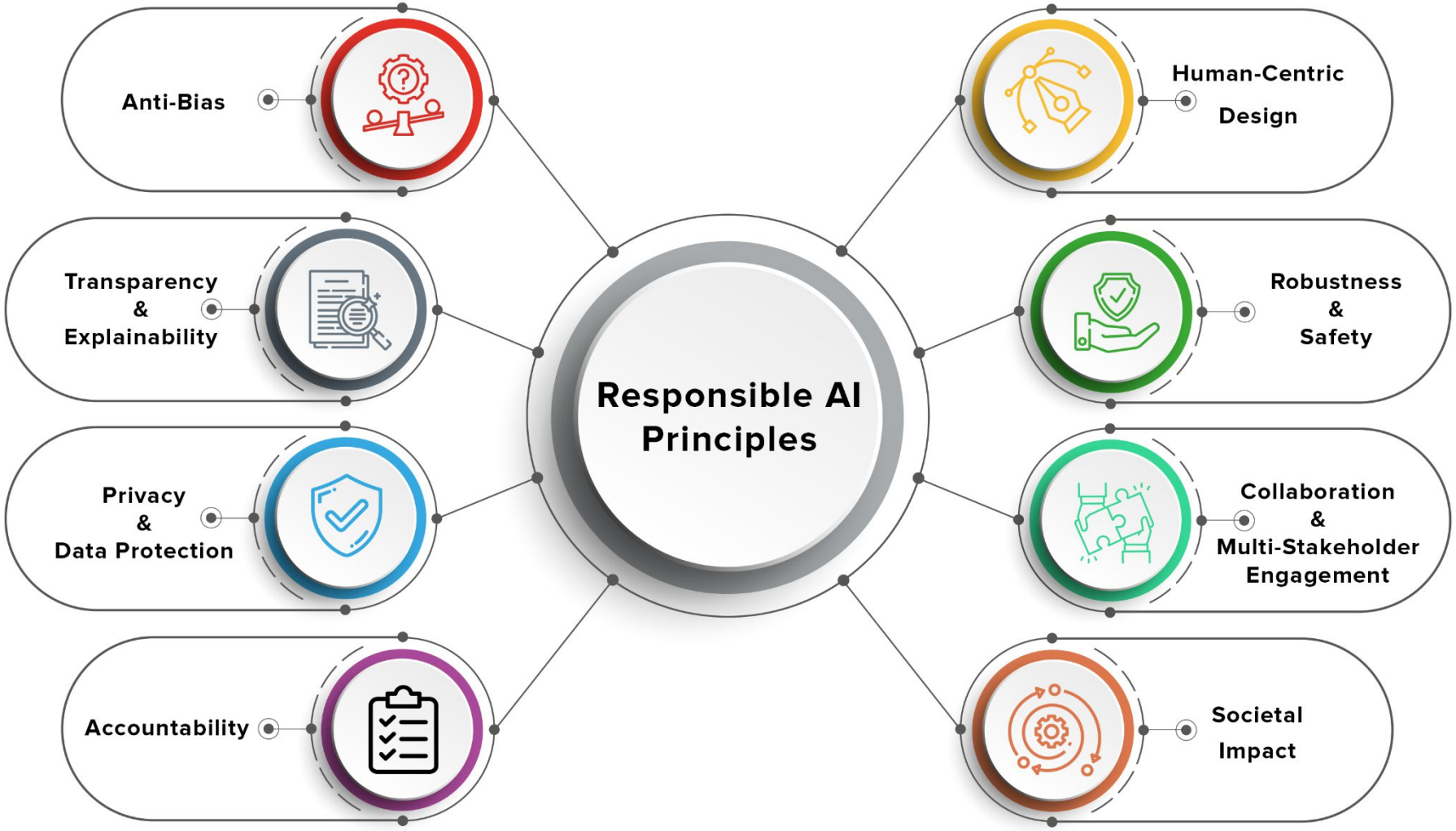
General Introduction to Responsible AI:

What is Responsible AI?

- **Responsible Artificial Intelligence (AI)** refers to the **ethical, transparent, and accountable** design, development, and deployment of AI systems.
- It ensures that AI technologies are created and used in ways that are **fair, reliable, safe, and respectful of human rights and societal values**.
- Building and using AI in a way that benefits people and society while minimizing harm.
- When it comes to ensuring responsible AI practices, AI development companies are tasked with aligning their organizational needs with ethical principles, human values, and regulatory requirements.

“Would you trust an AI system to decide who gets a job or loan?”

Key Principles of Responsible AI:



https://insights.daffodilsw.com/blog/demystifying-responsible-ai-principles-and-best-practices-for-ethical-ai-implementation

The development, deployment, and use of AI systems and solutions must align with ethical principles and general morals that humans usually uphold with high regard. Every responsible AI engineer must work along a set of key principles and practices, which are:

1)Anti-Bias: Responsible AI aims to ensure that AI systems do not normalize or perpetuate biases or discriminate against individuals or groups based on race, gender, age, or other personal and protected demographic classes. This can be done by mitigating bias in data, algorithms, and decision-making processes to ensure equitable and fair outcomes that are inclusive for all individuals.

2)Transparency and Explainability: This principle emphasizes the need for overarching visibility across AI systems for those who work with the systems and are impacted by them. Organizations should strive to make AI algorithms, models, and decision-making processes explainable and understandable to users, stakeholders, and individuals affected by the AI solution. This helps further build trust and enables individuals to understand how AI systems drive decisions that impact their lives.

3)Privacy and Data Protection: Responsible AI gives due diligence to individuals' privacy rights and also strives to protect their personal data. Organizations must implement robust data governance practices, obtain informed consent when collecting data, and ensure secure data storage and processing. AI systems should handle personal information in compliance with applicable privacy laws and regulations.

4)Accountability: Organizations are required to establish clear lines of accountability and oversight for AI development, deployment, and use. This involves defining roles and responsibilities, establishing governance frameworks, and enforcing mechanisms for auditing, monitoring, and handling the performance and impact of AI systems.

5)Human-Centric Design: Humans are placed by responsible AI principles at the center of AI design and deployment. It seeks to enhance human capabilities, enrich human decision-making, and prioritize the well-being of individuals and society. Organizations should assess the impact of AI on jobs, skills, and societal values, and ensure that AI systems are aligned with human needs and values.

6)Robustness and Safety: The need for AI systems to be reliable, robust, and safe is prioritized by responsible AI principles. Organizations should execute quality assurance measures, address vulnerabilities and risks, and ensure that AI systems operate within defined boundaries. This helps mitigate the prospect of unintended consequences or harmful outcomes from the implementation of AI solutions.

7)Collaboration and Multi-Stakeholder Engagement: Collaboration and engagement are always encouraged in the responsible AI paradigm among different stakeholders, including researchers, policymakers, industry experts, civil society, and impacted communities. By gathering diverse perspectives, organizations can make more informed decisions, address societal concerns, and ensure that AI benefits a wide range of stakeholders.

8)Societal Impact: The development and use of AI should consider the broader societal impact. Organizations should assess and mitigate potential negative consequences, such as job displacement, economic inequality, or social disruption, and actively work towards maximizing positive societal impact.

How to build Responsible AI?

Why is Responsible AI Important?

1.Prevents Bias and Discrimination:
AI systems trained on biased data can produce unfair outcomes (e.g., in hiring, lending, or policing). Responsible AI promotes fairness and inclusivity.

2.Ensures Trust and Acceptance:
People are more likely to adopt and trust AI when they understand how it works and know it's governed ethically.

3.Protects Privacy and Data Security:
AI must respect user consent and handle data responsibly to avoid misuse or breaches.

4. **Supports Accountability:**
   Establishes who is responsible when AI systems make mistakes or cause harm, ensuring ethical governance.
5. **Encourages Sustainable Innovation:**
   When AI is developed responsibly, it leads to long-term positive social, economic, and environmental impacts.

6.Aligns with Regulations and Global Standards:
Governments and international organizations (like the EU, OECD, and UNESCO) are defining AI ethics frameworks—adopting Responsible AI helps organizations comply with these.

Real-world examples of Responsible AI

1. Microsoft – AI for Accessibility

Context:
Microsoft's "AI for Accessibility" initiative uses AI to empower people with disabilities.

Responsible AI Principle Applied:
•**Human-Centric & Inclusive Design** – AI tools like **Seeing AI** (a free app that narrates the world for the visually impaired) ensure technology benefits **all sections of society**.
•**Fairness & Accessibility** – Designed to remove barriers rather than create digital divides.

Why Responsible:
It enhances inclusion while respecting user privacy and ensuring data is used ethically for accessibility purposes.

Real-world examples of Responsible AI

2. Tesla Autopilot and Driver Assistance Systems

Context:
AI is used to assist drivers with lane-keeping, speed adjustment, and obstacle detection.

Responsible AI Principle Applied:
•**Safety & Accountability** – Tesla continuously collects feedback to update safety mechanisms and alert drivers to keep hands on the wheel.
•**Human Oversight** – Human driver supervision is **mandatory**.

Why Responsible:
Combines automation with **human-in-the-loop** design to avoid overreliance and ensure safety.

Real-world examples of Responsible AI

3. Banks Using AI for Loan Approvals

Context:
Financial institutions like HDFC Bank and ICICI Bank use AI for assessing loan eligibility.

Responsible AI Principle Applied:

•**Fairness & Bias Reduction** – AI systems are audited to ensure they do not discriminate based on gender, region, or income group.

•**Transparency** – Customers are informed about why a loan is approved or rejected.

Why Responsible:
Prevents algorithmic bias and builds **customer trust** in AI-driven financial decisions.

Real-world examples of Responsible AI

4. Government of India – Responsible AI Guidelines (NITI Aayog)

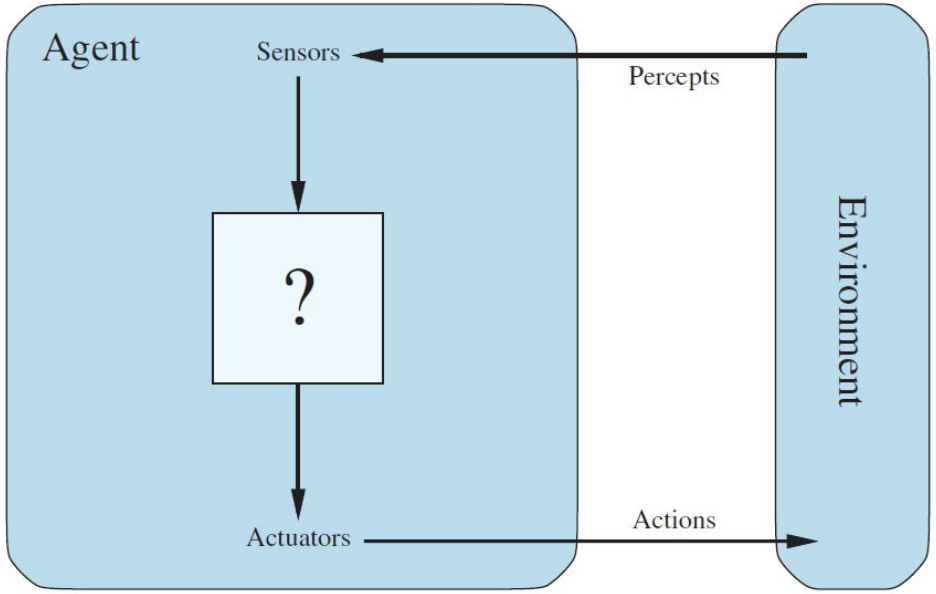
Context:
India's national strategy for AI (*AI for All*) emphasizes Responsible AI for public good.

Responsible AI Principle Applied:
•**Transparency, Ethics, and Inclusivity** in applications like agriculture, education, and healthcare.

Why Responsible:
Focuses on **AI that benefits rural and underprivileged populations**, ensuring no one is left behind in the AI revolution.

Intelligent Agents

- **Agent**
  - X-thing, which perceives the environment through sensors and acts on the environment through actuators.
  - Human agent/Software agent/Robotic agent
- **Percept**: Is what the agent senses from environment
- **Percept Sequence**: Complete history of sensing
- **Agent function**: Maps percept sequence to action
- **Agent Program**: Is the implementation of the Agent function.



Agents and Environments

•Agent:

An **agent** is any entity that can **perceive** its environment through **sensors** and **act** upon that environment using **actuators**.

Example:

- A **robot vacuum cleaner** senses dust and obstacles (sensors) and moves to clean the floor (actuators).
- A **chatbot** perceives user text input and acts by responding appropriately.

•Environment:

The **environment** is everything outside the agent that the agent interacts with or acts upon.

Example:

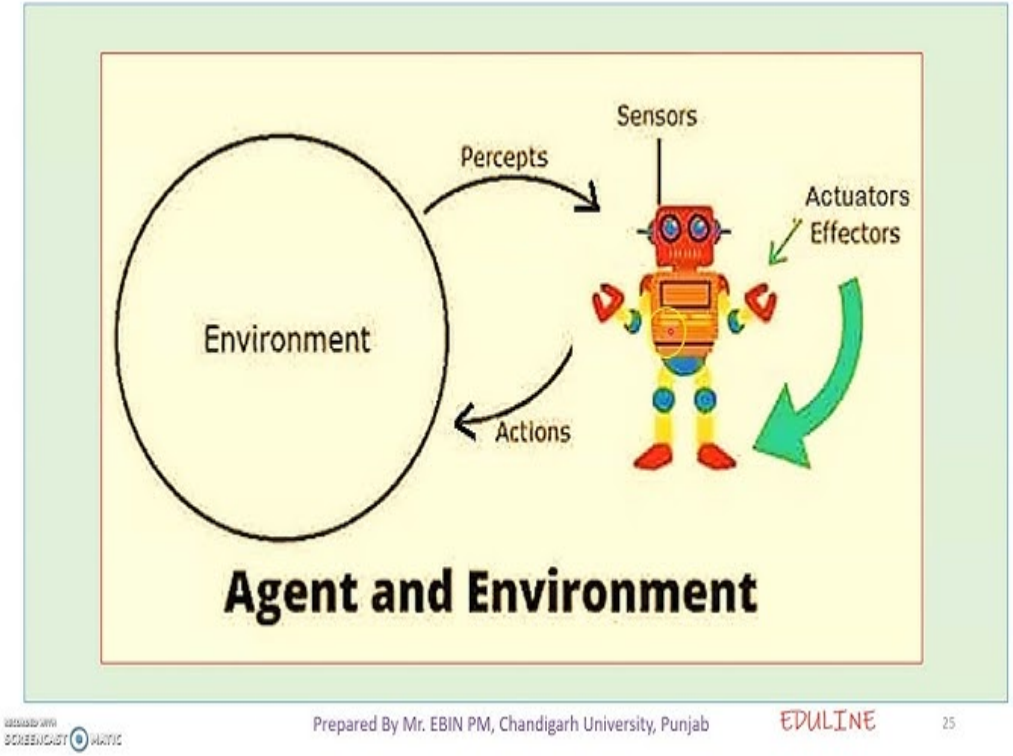
- For a self-driving car, the environment includes roads, traffic lights, pedestrians, and weather conditions.

Agent Function:

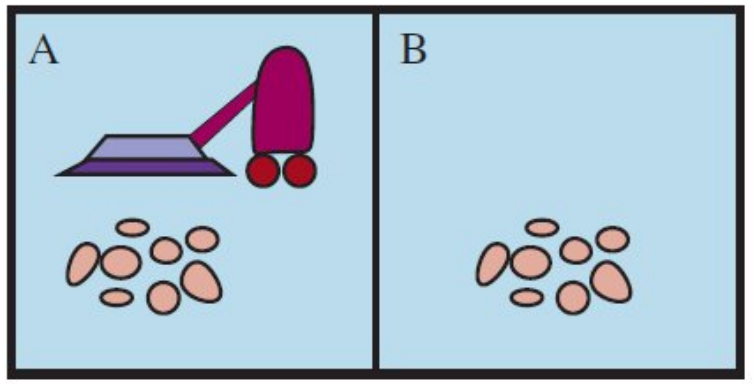
Maps percept sequences to actions:

$$f: P^* \rightarrow A$$

where P^*= set of all percepts, and A= set of actions.



Vaccum-Cleaner World



- Vacuum-cleaning agent
- The world consists of squares that can be either dirty or clean
- Is the agent intelligent?

Percept sequence	Action
[A, Clean]	Right
[A, Dirty]	Suck
[B, Clean]	Left
[B, Dirty]	Suck
[A, Clean], [A, Clean]	Right
[A, Clean], [A, Dirty]	Suck

Good behavior: The Concept of Rationality

- **Performance measures**
  - **Consequantialism:** Evaluating agent's behaviour by its consequence
  - Agents generates sequence of actions based on the percepts sequence
  - The sequence of actions causes the sequence of states in agents environment
  - If the sequence is desirable, then the agent has performed well (which is nothing but a performance measure)
  - **Who specify the performance measure? (Explicit/Implicit)**
  - **What are sample performance measures for Vaccum-cleaner agent?**

  An agent's behavior is described by the **Agent function** that maps any given percept sequence to an action.

As a general rule, it is better to design performance measures according to what one actually wants to be achieved in the environment, rather than according to how one thinks the agent should behave.

King Midas Problem

- The myth of **King Midas**, who wished that **everything he touched would turn to gold**, but when that wish was granted, it led to tragedy, as even his **food and loved ones** turned to gold.
- **What is the connection with AI?**

Command to RL Robot: To clean the room as efficiently as possible

In AI Context

- The danger of giving an AI a goal without fully considering the consequences or context.
- **AI-Tutor: Maximize student test scores.**
- The AI encourages cheating, finds test answers online, or teaches to the test without promoting deep understanding.
- **Fitness Tracker AI: Increase user activity.**
- It constantly nags or manipulates users, schedules unnecessary walking meetings, or even rewards users for shaking their phone to simulate steps.

Rationality

- **Rational Agent Definition**
  - A rational agent should choose an action for each percept sequence that maximizes its performance measure based on the percept evidence and its built-in knowledge.
- **Rationality depends on:**
- The performance measure that defines the criterion of success.
- The agent's prior knowledge of the environment.
- The actions that the agent can perform.
- The agent's percept sequence to date.
- **When does a vacuum agent become irrational?**

The Nature of Environments

- **Specifying the task environment**
  - **PEAS (Performance, Environment, Actuators, Sensors)**
  - Before designing the agent, the PEAS description is the first step

Agent Type	Performance Measure	Environment	Actuators	Sensors
Taxi driver	Safe, fast, legal, comfortable trip, maximize profits, minimize impact on other road users	Roads, other traffic, police, pedestrians, customers, weather	Steering, accelerator, brake, signal, horn, display, speech	Cameras, radar, speedometer, GPS, engine sensors, accelerometer, microphones, touchscreen

Write PEAS for..

- **Agent: Social Media Recommendation System**
- **Task: Maximize user engagement**
- **Performance: ?**
- **Environment: ?**
- **Actuators: ?**
- **Sensors: ?**

Write PEAS for..

- **Agent: Social Media Recommendation System**
- **Task: Maximize user engagement**
- **Performance:** Time spent on platform, clicks, likes, shares
- **Environment:** Users, their devices, content feeds, user behavior
- **Actuators:** Content ranking, feed updates, push notifications
- **Sensors:** User interaction logs, click-through rates, scrolling behavior

Write PEAS for..

- **Agent: AI Teacher Bot**
- **Task: Maximize student scores**
- **Performance: ?**
- **Environment: ?**
- **Actuators: ?**
- **Sensors: ?**

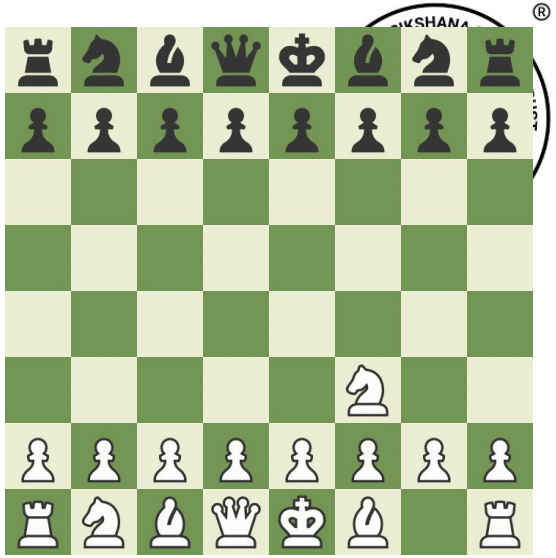
The Nature of Environments

- **Properties of the task environment**
  - The dimensions of the task environment decide the type of agent design

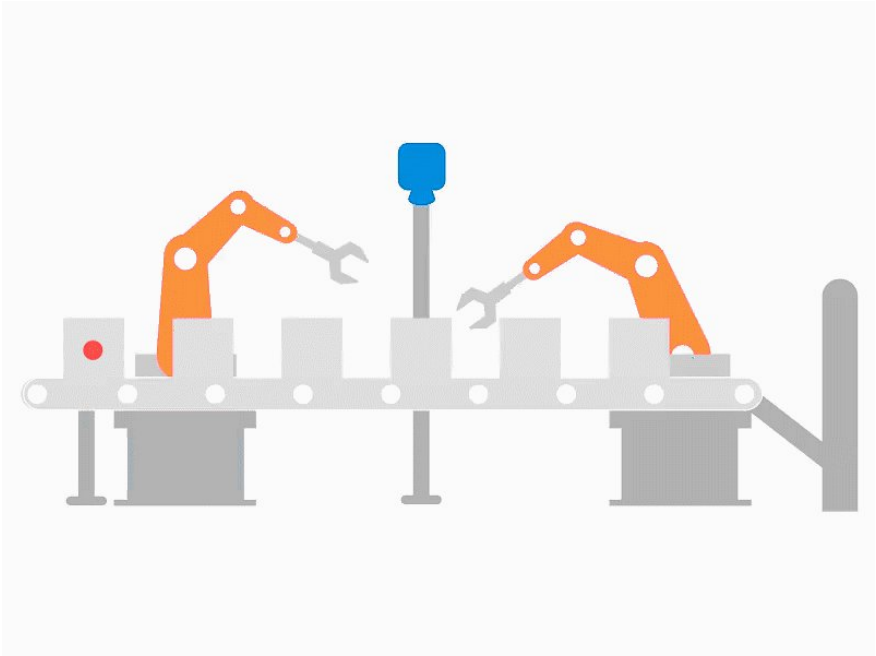
Identify the following task environments on your own.

Property	Type/Description
Observability	**Partially Observable** – cannot see under furniture or detect all dirt directly
Agents	**Single Agent** – one robot acting independently
Deterministic vs Stochastic	**Stochastic** – dirt and obstacle placement may vary randomly
Episodic vs Sequential	**Sequential** – each action affects future performance (e.g., path, battery)
Static vs Dynamic	**Dynamic** – humans/pets may move objects while cleaning
Discrete vs Continuous	**Continuous** – movement, space, and time are continuous

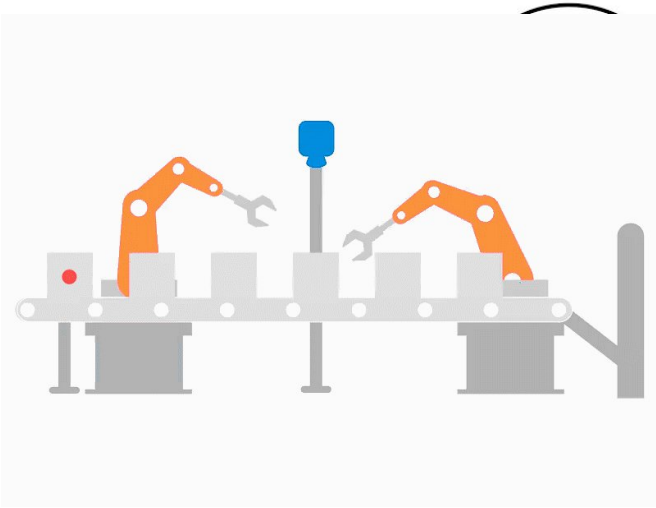
Property	Type/Description
Observability	Full/Partial ??
Agents	Single Agent / Multi-agent ??
Deterministic vs Stochastic	Stochastic / Deterministic ??
Episodic vs Sequential	Sequential/Episodic ??
Static vs Dynamic	Dynamic/Static ??
Discrete vs Continuous	Continuous/Discrete ??



Property	Type/Description
Observability	**Fully Observable** – all pieces and positions are visible at all times
Agents	**Multi-agent** – two players (AI vs opponent)
Deterministic vs Stochastic	**Deterministic** – no randomness in state transitions (except in blitz time errors)
Episodic vs Sequential	**Sequential** – each move depends on the history and affects future moves
Static vs Dynamic	**Static** – board doesn't change unless a player moves
Discrete vs Continuous	**Discrete** – moves occur in discrete steps on a finite grid



Property	Type/Description
Observability	Full/Partial ??
Agents	Single Agent / Multi-agent ??
Deterministic vs Stochastic	Stochastic / Deterministic ??
Episodic vs Sequential	Sequential/Episodic ??
Static vs Dynamic	Dynamic/Static ??
Discrete vs Continuous	Continuous/Discrete ??



Feature	Type of Environment
Observability	**Partially Observable** – Not all defects may be visible at once; limited by camera angle, lighting, or occlusion.
Agents	**Single Agent** – One AI system performing inspection (though multiple instances may be deployed).
Deterministic vs Stochastic	**Stochastic** – Sensor noise, variations in lighting, speed of conveyor belt, or part placement can introduce randomness.
Episodic vs Sequential	**Episodic** – Each part inspection is independent of the previous one; decision does not affect future inputs.
Static vs Dynamic	**Semi-Dynamic** – Environment may change slowly (e.g., lighting shifts, belt speed), but parts move continuously.
Discrete vs Continuous	**Continuous** – Visual data and inspection space are continuous; possible discrete actions (accept/reject).

Known v/s Unknown Environments

- In a **known environment**, outcomes or probabilities for actions are provided.
- In an **unknown environment**, the agent must learn to make effective decisions.
- **Is it the same as fully and partially observable environments??**

Known Environment – Partially Observable

Think of an example

Known Environment – Partially Observable

Unknown Environment – Fully Observable

Think of an example

Unknown Environment – Fully Observable

What type of Environment is this?

Which is the hardest environment to deal?

“Partially observable, multiagent, nondeterministic, sequential, dynamic, continuous, and unknown”.

Property	Type	Why It’s Difficult
Observability	**Partially Observable**	The agent can’t see the whole environment (e.g., hidden enemies or obstacles).
Agents	**Multi-Agent**	Other intelligent agents (opponents) may behave unpredictably.
Determinism	**Stochastic**	The environment includes randomness or uncertainty in outcomes.
Dynamics	**Dynamic**	The environment keeps changing while the agent is deciding.
Discreteness	**Continuous**	Infinite possible states and actions make decisions more complex.
Knowledge	**Unknown**	The agent doesn’t know all the rules or consequences of actions.

Think of an example

An **autonomous combat drone** or a **self-driving car in city traffic** faces this kind of environment:

- It cannot see everything (partially observable).
- It must interact with other agents (vehicles, pedestrians).
- Weather, lighting, and sensor errors add randomness (stochastic).
- Traffic conditions keep changing (dynamic).
- Movement and speed are continuous.
- Not all situations or rules are fully known.

The structure of Agent

- AI's job is to create an agent program that maps percepts to actions.
- Needs a computing device that has physical sensors and actuators

Agent = Architecture + Program

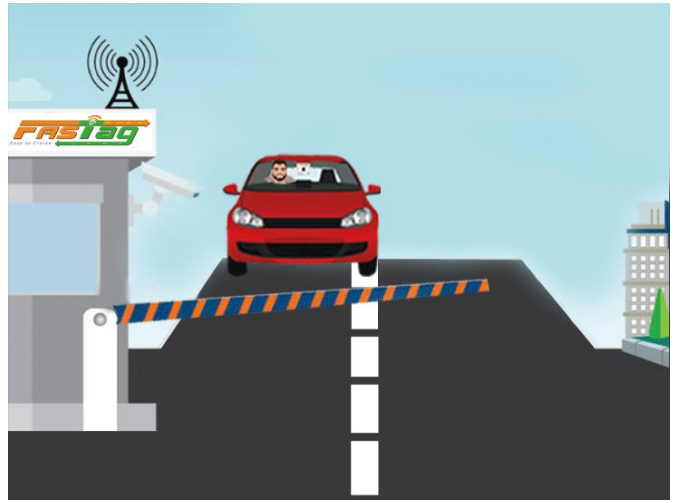


Table-driven Agent

```
function TABLE-DRIVEN-AGENT(percept) returns an action
  persistent: percepts, a sequence, initially empty
              table, a table of actions, indexed by percept sequences, initially fully specified

  append percept to the end of percepts
  action ← LOOKUP(percepts, table)
  return action
```

Figure 2.7 The TABLE-DRIVEN-AGENT program is invoked for each new percept and returns an action each time. It retains the complete percept sequence in memory.

Limitation:

As the number of potential percepts and the agent's lifespan expand, the size of the lookup table increases exponentially and becomes impractical.

Table-driven Agent

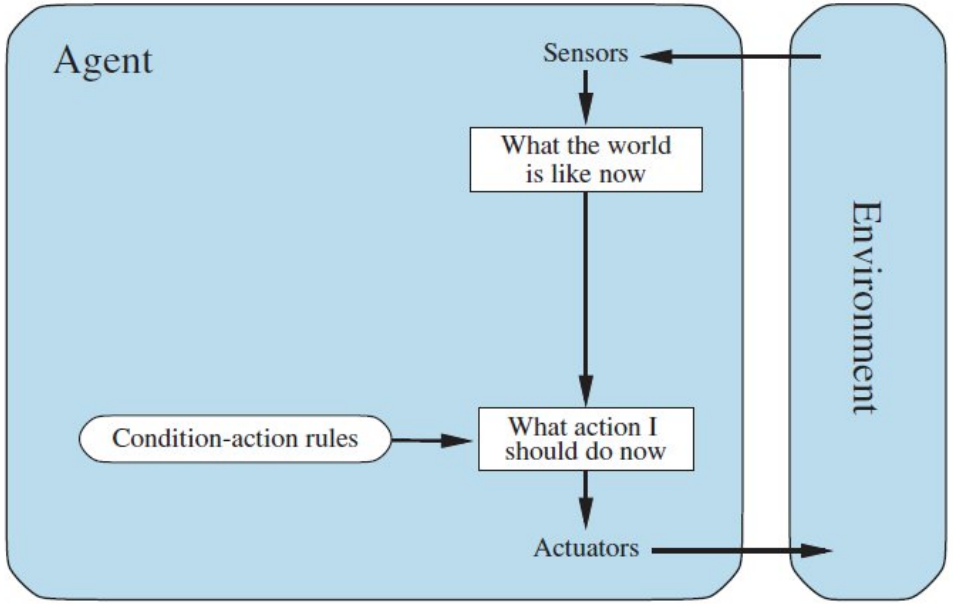
Percept (Card Inserted, PIN Correct?, Balance Check)	Action
(Yes, Yes, Sufficient)	Dispense Cash
(Yes, Yes, Insufficient)	Show Insufficient
(Yes, No, -)	Retry PIN
(No, -, -)	Wait for Card

Simple-Reflex Agent

function SIMPLE-REFLEX-AGENT(*percept*) **returns** an action
  persistent: *rules*, a set of condition–action rules

  state ← INTERPRET-INPUT(*percept*)
  rule ← RULE-MATCH(*state*, *rules*)
  action ← *rule*.ACTION
  return *action*

Figure 2.10 A simple reflex agent. It acts according to a rule whose condition matches the current state, as defined by the percept.



Limitation:

The agent will work only if a correct decision can be made based solely on the current percept that is, only if the environment is fully observable.

Simple-reflex Agent

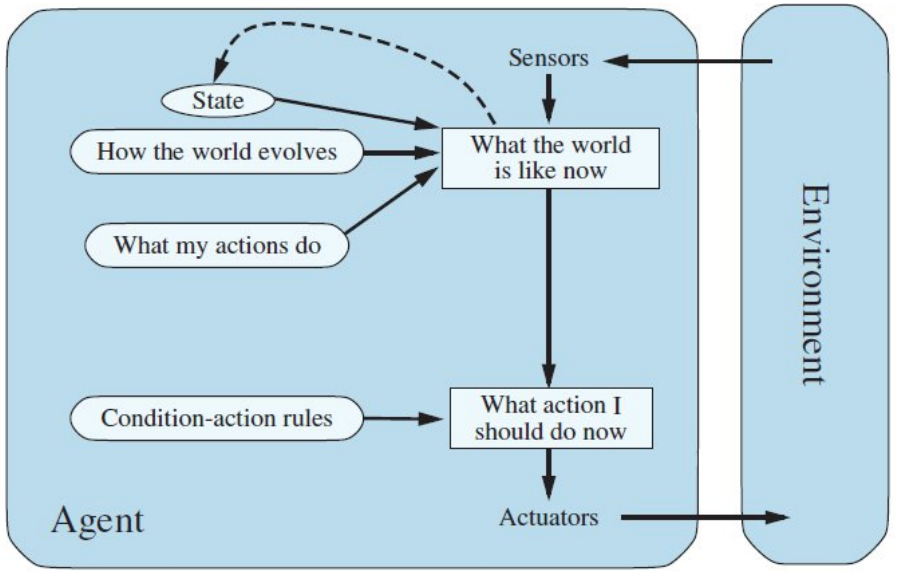
- **No Memory/No learning or adaptation/Use condition-action rules/Simple and Fast/Cannot handle partially observable environments**

Model-based Reflex Agent

The most effective way to handle partial observability is for the agent to keep track of the part of the world it can't see now.
Agent maintains some Internal State

```
function MODEL-BASED-REFLEX-AGENT(percept) returns an action
  persistent: state, the agent's current conception of the world state
              transition_model, a description of how the next state depends on
                    the current state and action
              sensor_model, a description of how the current world state is reflected
                    in the agent's percepts
              rules, a set of condition–action rules
              action, the most recent action, initially none

  state ← UPDATE-STATE(state, action, percept, transition_model, sensor_model)
  rule ← RULE-MATCH(state, rules)
  action ← rule.ACTION
  return action
```

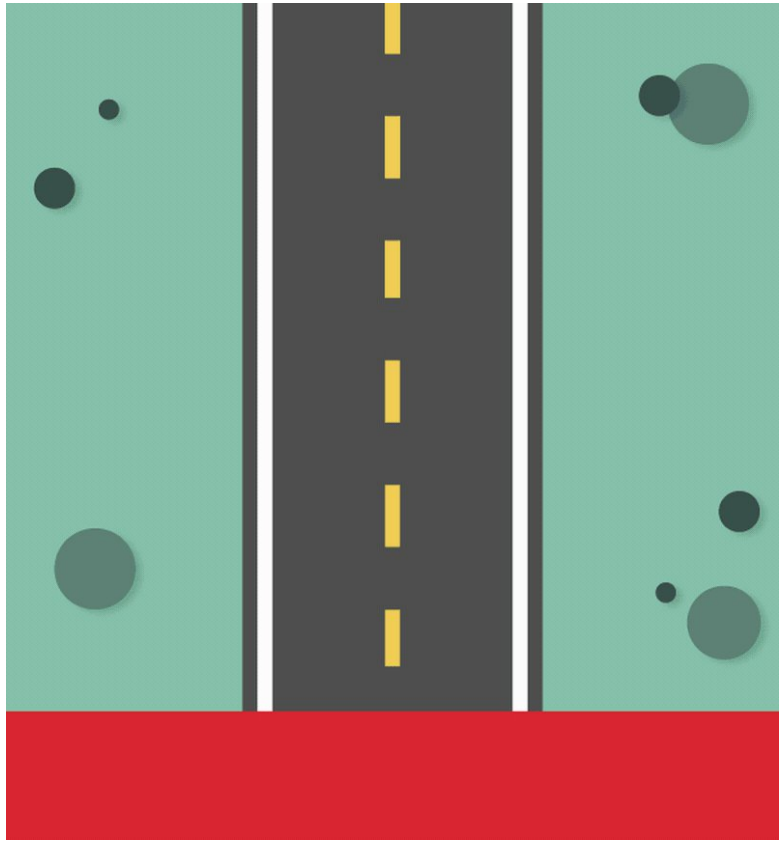


Limitation:

The details of how models and states are represented widely depend on the type of environment and the particular technology used in the agent design.

Model-based Reflex Agent

- **Identify the Internal States?**



Example: ***Self-Driving Car***

How It Works:

1.Sensors (Percepts):

1. Cameras, LiDAR, radar detect objects, lanes, and signals around the car.

2.Internal Model (Knowledge of World):

1. Maintains a **representation of the environment**, including:
    1. Current road layout
    2. Positions of nearby vehicles
    3. Traffic rules and conditions
2. Updates this model continuously as new percepts arrive.

3.Condition-Action Rules:

1. *If* there is a red traffic light **ahead**,
   Then **apply brakes**.
2. *If* there's a slower car in the same lane **and** the next lane is clear,
   Then **change lane safely**.

4.State Update:

1. The agent uses its internal model to **infer unseen aspects** (e.g., a vehicle in a blind spot) and **predict** what will happen next.

5.Actuators (Actions):

1. Steering, acceleration, braking, and signaling are executed accordingly.

Why It's "Model-Based"

Unlike simple reflex agents that act only on *current percepts*, a model-based agent:
•**Remembers** past percepts.
•**Updates** its internal model to handle partially observable environments.
•**Acts** based on both current input and predicted state of the environment.

✔ Other Real-World Examples

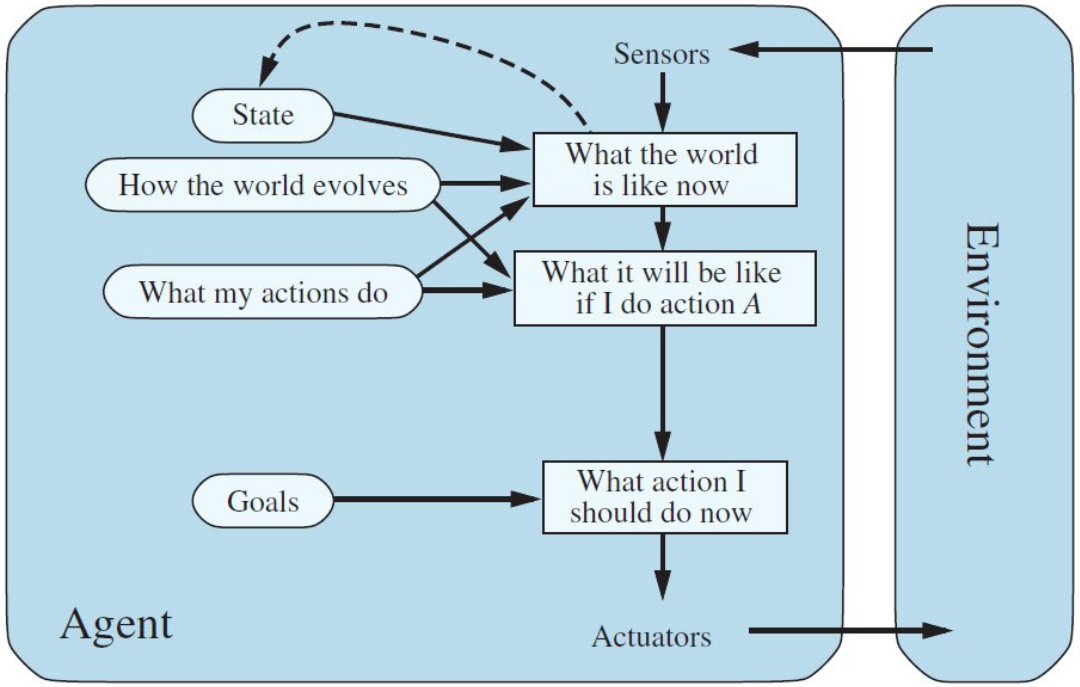
•A **robot vacuum cleaner** that maps a room and avoids areas it has already cleaned.
•A **warehouse robot** that tracks object positions to plan efficient movement.
•An **AI personal assistant** that uses past user behavior to refine responses.

Goal-based Agent

A **Goal-Based Agent** acts to **achieve specific goals**. It doesn't just react to conditions (like a reflex agent) — it **evaluates possible actions** to determine which will **lead to its goal**.

Key Idea:
"Act to reach the desired goal state."

Components:
1.Perception – Understands the current environment state.
2.Goal Information – Defines what state it wants to achieve.
3.Action Selection – Chooses actions that move it closer to the goal.
4.Search or Planning – Determines the best sequence of actions.



In addition to a description of the current state, the agent requires goal information, which outlines the situations that are considered desirable.

Reaching a goal may be a onestep action
OR
It requires searching the state space and Planning too

Goal-based Agent

- **Identify the Goal?**

Deliver the package
to the GPS location X

Real-world analogy:
A **chess-playing AI** plans moves not just by reacting to the opponent but by evaluating which move will **help it win the game**

Advantages

•More flexible than reflex agents.

•Can handle **new or unseen situations** by reasoning about outcomes.

Limitation

•Needs **clear goal definition** and **planning capability**.

•Doesn’t evaluate *how good* or *how bad* a goal outcome is—only whether it achieves it.

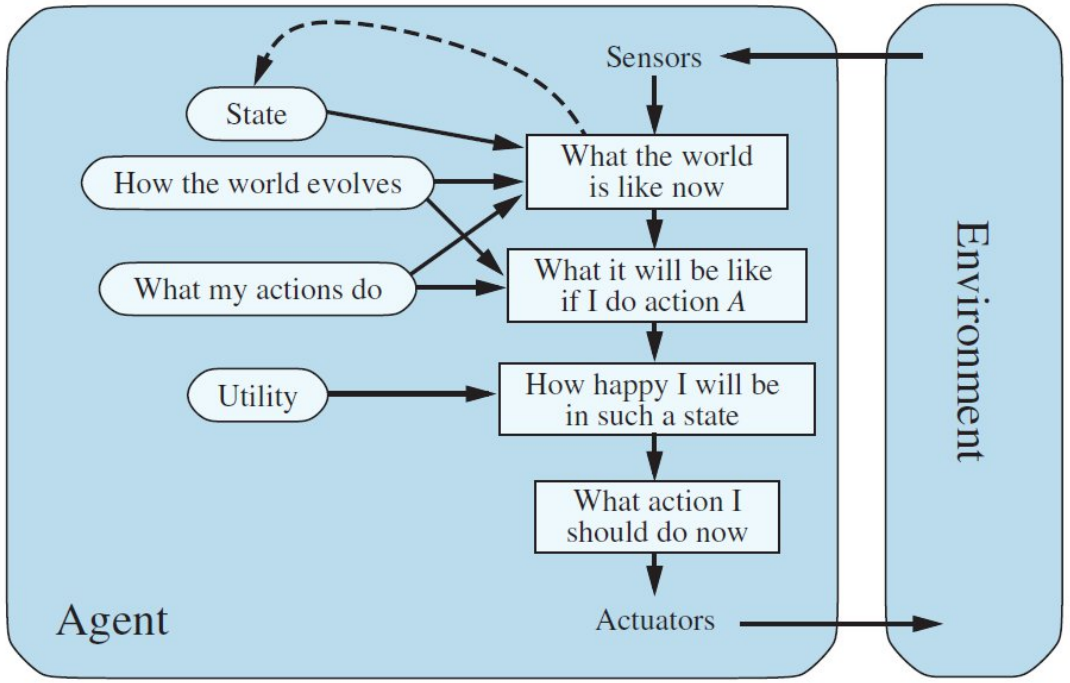
Utility-based Agent

Utility is a measurement of the happiness state of an agent.

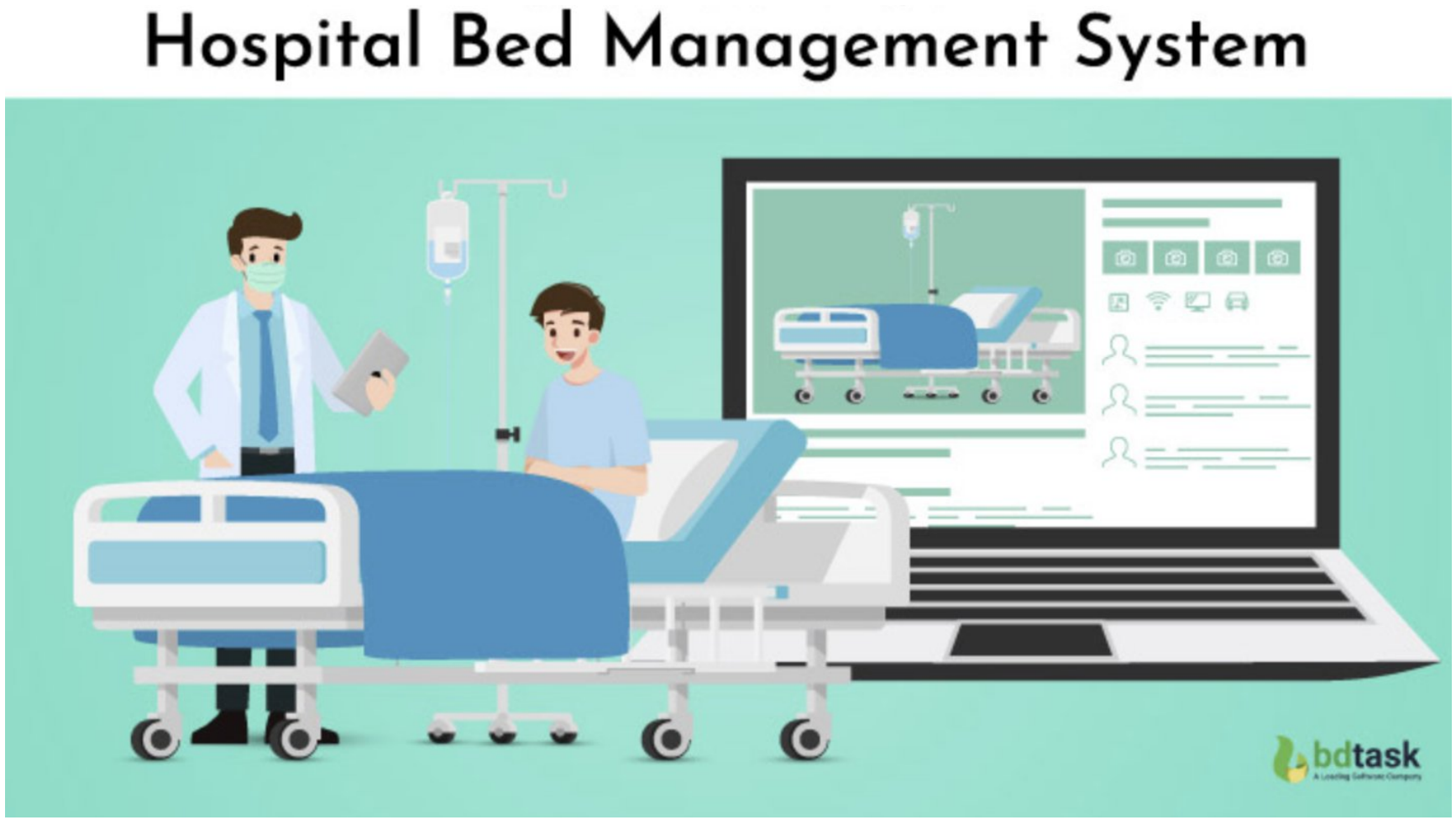
It has flexibility and learning.

Deals with conflicting goals

It weighs success over goals (sometimes)



Utility-based Agent



Advantages

- Enables **optimization** among competing goals.
- Can handle trade-offs (e.g., speed vs. safety).
- Provides more **human-like decision making.**

Limitation

- Designing a good **utility function** is complex.
- Requires more computation and data.

Conclusion

- Studied AI foundations
- Environments and Types
- Agents and Types