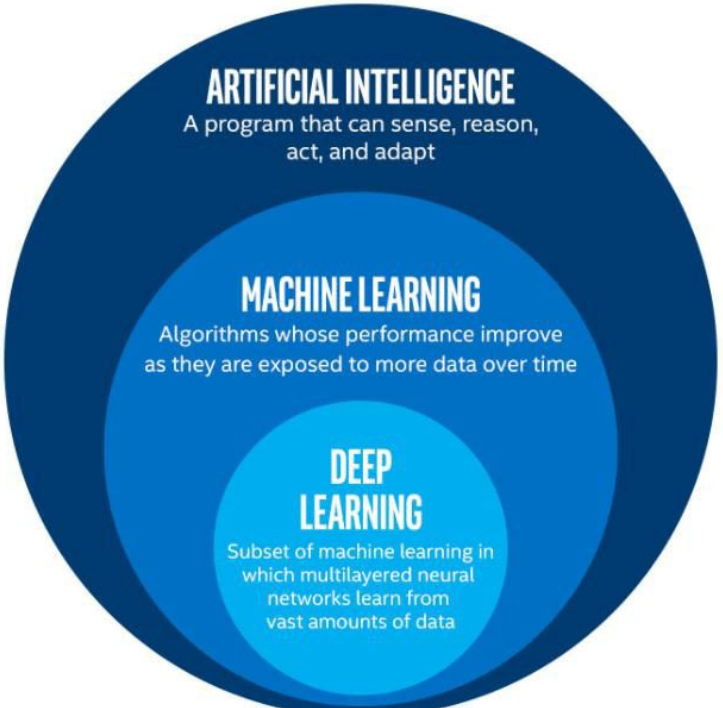
ICS: 2308

Artificial Intelligence

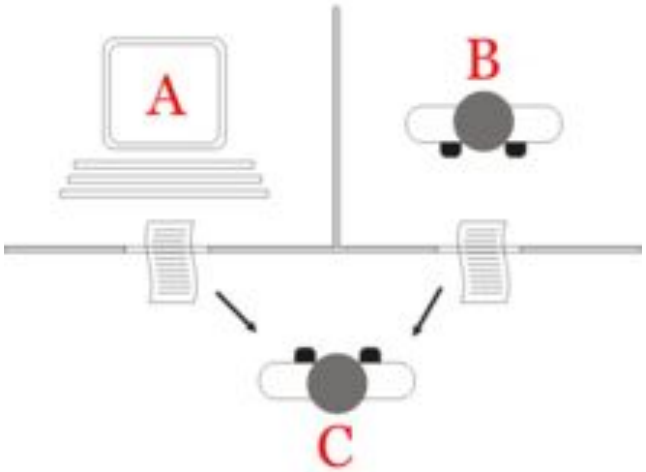
Can We Make Intelligent Machines ?

Human performance vs Rationality

Thought vs Behavior



Acting humanly: The Turing test approach



Thinking humanly: The cognitive modeling approach

How we learn about our brains.

- Introspection
- Psychological experiments
- Brain imaging

1.1.3 Thinking rationally: The “laws of thought” approach.

Aristotle's **syllogisms** provided patterns for argument structures that always yield correct conclusions when given correct premises.

Field of Logic.

By 1965, programs existed that could, in principle, solve any solvable problem described in logical notation.

The **logicist tradition** hopes to build on such programs to create intelligent systems.

1.1.3 Thinking rationally: The “laws of thought” approach.

Obstacles to the approach:

a) It requires knowledge of the world that is certain, which in reality is seldom achieved.

- Theory of probability:

b) Lack of Intelligent Behavior from Thought Alone - Rational thinking, by itself, does not necessarily result in intelligent or useful actions.

- Theory of rational action

1.1.4 Acting rationally: The rational agent approach

Agent - Something that acts.

Rational agent is one that acts so as to achieve the best outcome or, when there is uncertainty, the best expected outcome.

To act rationally, an agent **may** require:

1. Correct inferences from the ‘thinking rationally” approach, (Though this is not all of rationality)
   - There may be no provably correct thing to do but something must still be done
   - Some reflex actions are more successful than slower ones taken after rational thought
2. The skills needed for a Turing test allow and agent to act rationally.

1.1.4 Acting rationally: The rational agent approach

Advantages over other approaches:

1. It is more general than the “laws of thought” approach - correct inference is just one of several possible mechanisms for achieving rationality.
2. It is more amenable to scientific development than approaches based on human behaviour and thought.

*AI has focused on the study and construction of agents that **do the right thing***.What counts as the right thing is defined by the objective that we provide to the agent

This general paradigm can be said to be the **Standard Model.**

Perfect rationality is not feasible in complex environments. The computational demands are too high.

1.1.5 Beneficial machines

The standard model is useful but probably not the right model in the long run - the standard model **assumes that we will supply a fully specified objective to the machine.**

The value alignment problem - The problem of achieving agreement between our true preferences and the objective we put into the machine. This matters because:

1. Incorrectly specified objectives in capable intelligent systems that are deployed in the real world can cause negative consequences
2. Intelligent systems might misbehave in unexpected ways like exploiting loopholes to achieve goals e.g bribing an opponent to win a chess match.

1.1.5 Beneficial machines

We don’t want machines that are intelligent in the sense of pursuing their objectives; we want them to pursue our objectives.

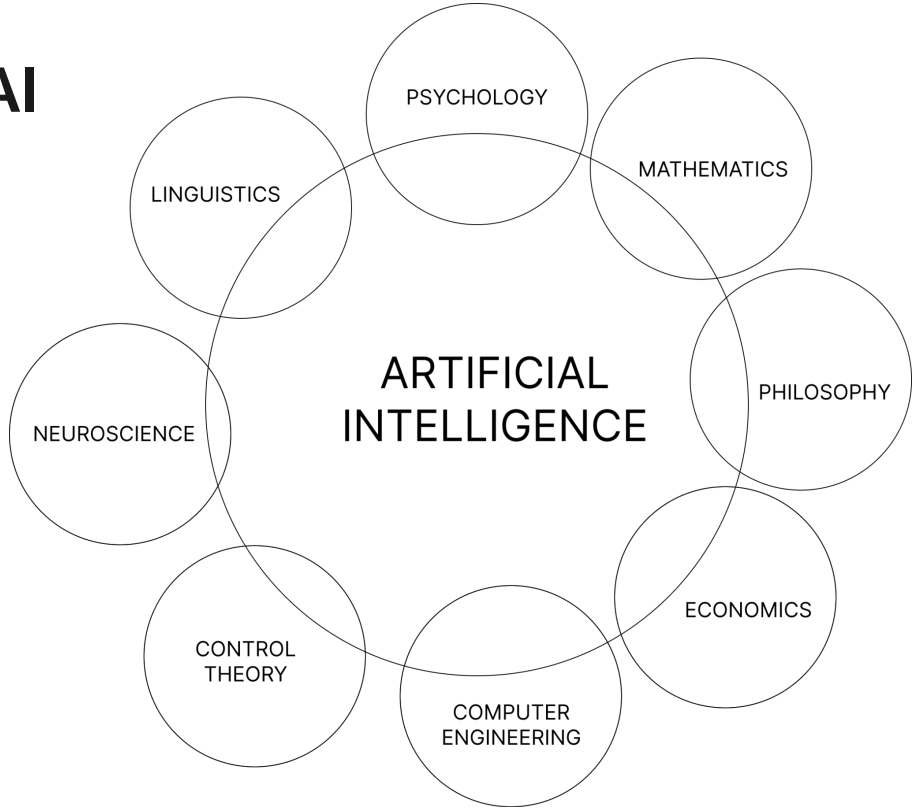
If we cannot transfer those objectives perfectly to the machine, then we need a new formulation

1. Uncertainty Awareness: Machines should recognize they don’t fully know human preferences.
2. Cautious Behavior: Acting cautiously and asking for permission when uncertain.
3. Learning from Observation: Continuously refining understanding of human objectives.
4. Deferring to Human Control: Prioritizing human oversight and input.

Goal: To have **provably beneficial** by pursuing human-aligned objectives instead of rigid, predefined goals.

Disciplines that make up AI

AI is a complex subject that is made up of various disciplines. We will focus on 8 disciplines how they contributed ideas, viewpoints and techniques towards AI.



Philosophy

Philosophy as a study helps us understand how humans think, where knowledge comes from and how humans use knowledge to formulate actions.

Scholars have tried to build systems that emulate how the mind operates according to logic/numerical rules. For instance, Aristotle came up with syllogisms.

The mind is more than just a physical entity

The problem with emulating the mind solely as a physical entity is that the mind is not entirely governed by physical laws. Rene Descartes argues that the mind consists of an aspect of the soul/spirit i.e. Dualism.

In contrast to Dualism, Materialism states that the brain operates according to laws of physics and with time the mind matures as a product of the brain's physical functions.

Source of knowledge

Empiricism states that knowledge acquisition is grounded on senses.

Induction. Knowledge is acquired by exposure to repeated associations between the general rules of interest.

Logical Positivism. Knowledge is grounded in logical theories and statements and these statements need to be empirically verifiable or analytically true.

Confirmation Theory. Acquisition of knowledge from experience by quantifying the degree of belief.

From knowledge to action

Intelligence is a combination of thinking and acting rationally based on the available knowledge.

Actions are a result of logical connections between goals and knowledge of their outcomes. This means following logical steps towards a goal.

Another approach involves not focusing on the goal itself, but how to achieve it. From the goal, we track backwards till the very first possible step and finally act towards the goal following the tracked steps.

Mathematical

The transition from a philosophical view of AI to a formal science requires the mathematization of logic and probability.

A formal rule is a set of rules that help us determine whether a statement is true or false. They are divided into:

1. Boolean Logic -> true, false, AND, OR and NOT
2. First Order Logic -> extends boolean logic allowing reasoning of object, their attributes and their relationships.

What are the limits of computation?

Not everything can be computed. This is because some problems are impossible and/or are too complex to solve.

Tractability. This is the solvability of a problem. A problem is tractable if it can be solved within reasonable time. Inversely, a problem is **intractable** if it takes too long to solve it.

Bridging the gap of uncertainty with probability and statistics

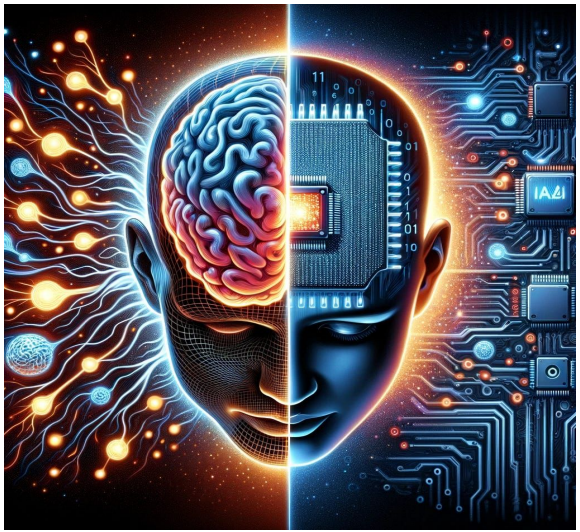
Probability can be seen as generalizing logic to situation with uncertain information. Statistics is a superset of probability in that it entails probability combined with the availability of data.

Sometimes, AI has inadequate information to solve a problem or even make a decision. An educated guess based on the available data is therefore made allowing AI to give a solution that is likely to be accurate.

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?

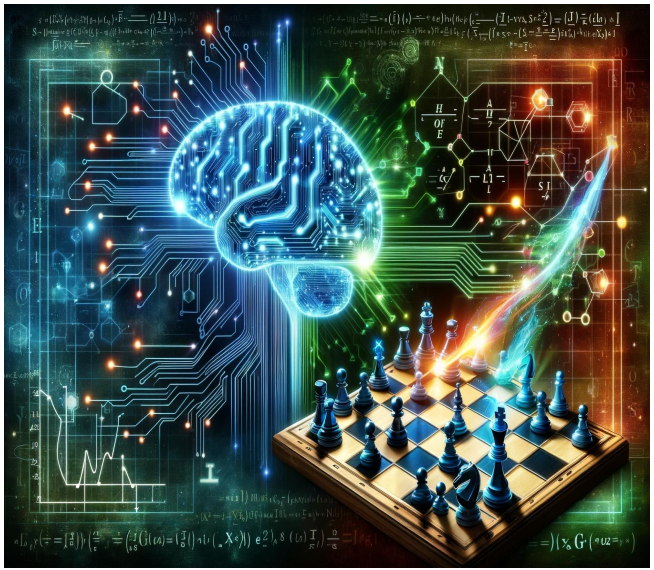
Decision Theory (Rational Choice Under Uncertainty)



- Combines **probability theory** with **utility theory**
- Provide a formal and complete framework for individual decisions made under uncertainty
- Uses **expected utility maximization** to guide rational agents.
- Works well in **large economies**, where individual actions have little effect on others.
- Forms the basis for AI models such as **Bayesian networks** and **reinforcement learning**.

Game Theory (Strategic Decision-Making)

- Developed by **von Neumann & Morgenstern (1944)**.
- Rational agents should adapt policy that are *(or at least appear to be)* randomised
- Essential in **multi-agent AI systems**, where actions influence other agents.



Operations Research (Sequential Decision-Making)

- Emerged in World War II from efforts in Britain to optimize radar installations, and later found innumerable civilian applications.
- Studies **long-term payoffs**, beyond just immediate decisions.
- ***Richard Bellman*** (1957) introduced **Markov Decision Processes (MDPs)**.
- MDPs are fundamental in **reinforcement learning**, **path planning**, and **robotics**.

Satisficing (Human-Like Decision-Making)

- Proposed by **Herbert Simon (1947),(1978 He won Nobel Prize)**.
- Making decisions that are “good enough” rather than laboriously calculating an optimal decision.
- Real-world agents (including humans) often use **heuristics (**mental shortcuts that people use to make decisions and solve problems**)** instead of full optimization.

Neuroscience

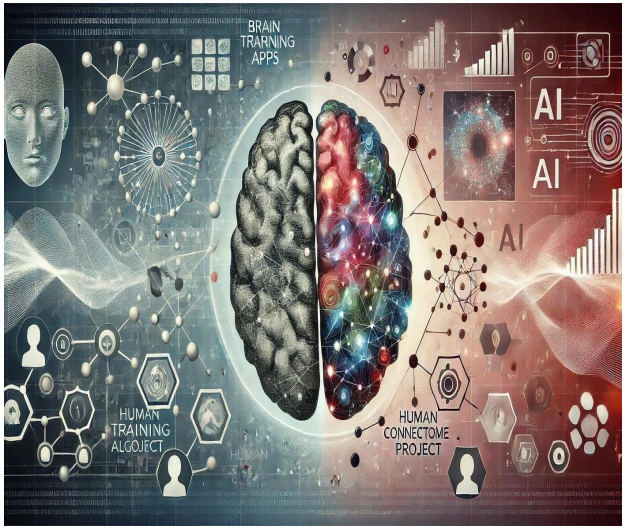
- Neuroscience plays a crucial role in AI because the human brain is the **most advanced known intelligence system**.

How does the brain process information?

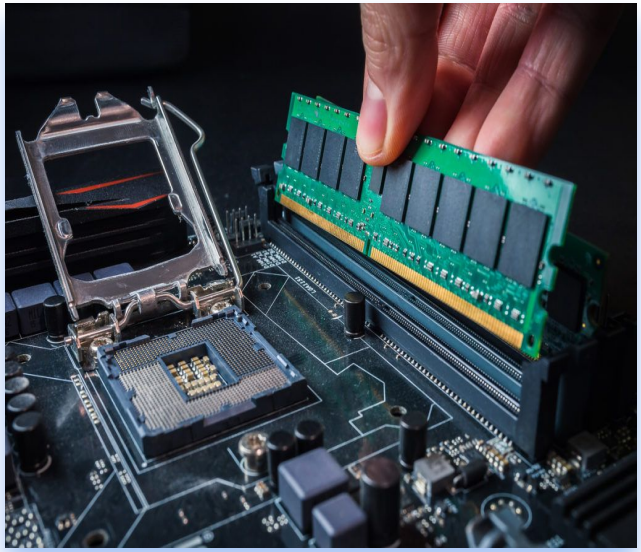
- The brain is composed of **neurons** that transmit signals through electrical and chemical processes.
- AI takes inspiration from this to create **artificial neural networks (ANNs)**.

How do neurons help in decision-making and learning?

- The brain strengthens connections between neurons when learning, a process called **synaptic plasticity**.
- In AI, this idea is mimicked using **backpropagation** and **weight updates** in neural networks.



Computer Engineering



Computer engineering provides the hardware and system-level optimizations required to efficiently run AI models. It focuses on designing and improving the physical and architectural foundations of AI systems.

How can we build an efficient computer?

- AI requires ***high computational power***, which has led to the development of specialized processors.
  i. GPUs *(Graphics Processing Units)*: Essential for parallel processing in deep learning.
  ii. TPUs *(Tensor Processing Units)*: Optimized for machine learning workloads.
  iii. WSE *(wafer scale engine)*: Optimized to train large neural networks significantly faster than GPUs.

Psychology

Psychology studies how people **perceive**, **reason**, and **make decisions**

How do humans and animals think and act?

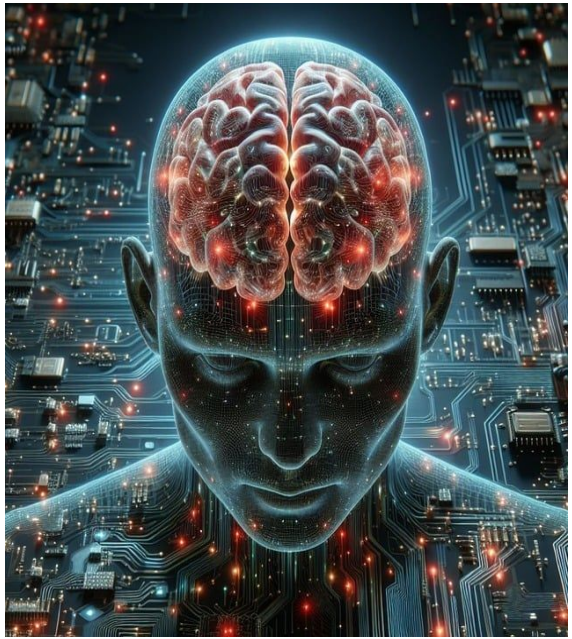
➢ AI applies concepts from psychology to **machine learning and problem-solving models**.

How do humans learn from experience?

➢ AI mimics this using **supervised learning, reinforcement learning, and neural networks**.

Can we replicate human cognitive abilities in AI?

➢ AI models like **expert systems, neural networks, and cognitive architectures** are inspired by human cognition.

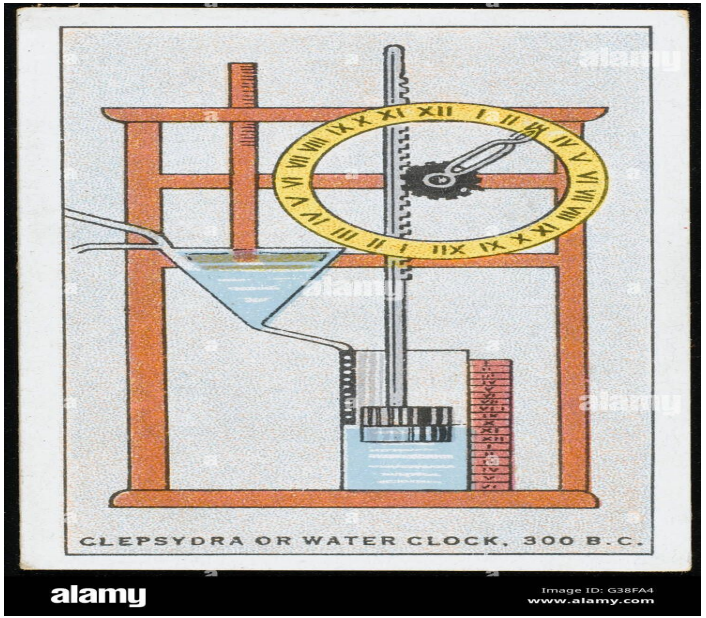


Control Theory and Cyberkinetics

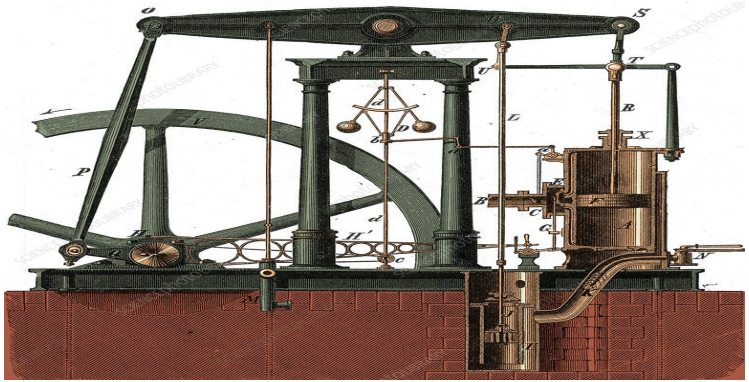
How can artifacts operate under their own control?

- Artifacts that operate autonomously rely on **control systems** that adjust their behavior in response to environmental changes.

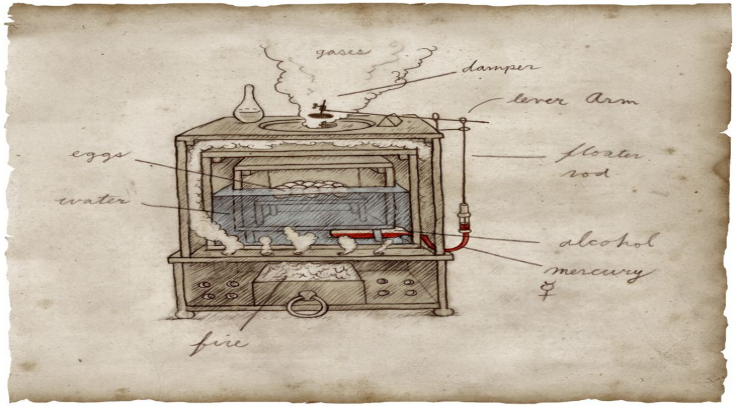
- The first known self regulating artifact, **Ktesibios' Water Clock (c. 250 BCE)**



Control Theory and Cyberkinetics (Cont)



James Watt's Steam Engine Governor (1736–1819): A centrifugal regulator that maintained a constant engine speed.



Cornelis Drebbel's Thermostat (1572–1633): A mechanism that automatically controlled temperature, demonstrating early homeostasis in machines.

Linguistics

How does language relate to thought?

- Understanding language requires an understanding of the subject matter and context, not just an understanding of the structure of sentences.

- **Knowledge representation** (the study of how to put knowledge into a form that a computer can reason with) was tied to language and informed by research in linguistics, which was connected in turn to decades of work on the philosophical analysis of language.

- AI needs more than words; it needs context, logic, and common sense to understand meaning.