Lecture 1 - Chpt 1, Chpt 2

Harley Caham Combest

Fa2025 CS4013 Lecture Notes – Mk1

. . . . . .

Chapter 1: Introduction to Artificial Intelligence

. . . . . .

Historical Context. Humans have long asked how the mind works. Aristotle studied reasoning rules. Alan Turing reframed the problem in 1950: instead of asking "Can machines think?" he asked if a machine could *imitate* human conversation well enough to fool an interrogator. This thought experiment became the **Turing Test**.

Definitions of AI. Approaches differ by whether they emphasize **thinking** vs. **acting**, and **human-like** vs. **rational**. Four traditions emerge:

Thinking Humanly	Thinking Rationally
Acting Humanly	Acting Rationally

Definition 1 (AI). Artificial Intelligence is the field concerned with building systems that display behavior we would call "intelligent" if done by humans.

Core Capabilities. Typical ingredients of an AI system:

- Natural language processing (understanding words).
- Knowledge representation (storing facts).
- Automated reasoning (drawing conclusions).
- Machine learning (improving with experience).
- Computer vision (seeing).
- Robotics (acting physically).

Calculative Example: The Turing Test. Setup. Imagine a human judge converses (via text) with both a machine and a person. If after many questions the judge cannot reliably tell which is the human, the machine *passes*.

- Machine must manage small talk, factual questions, and follow-ups.

- Requires integration of *all six AI capabilities*.

Walkthrough:

1. Judge: "What is 2+2?" Machine: "4." (trivial).
2. Judge: "Tell me a joke." Machine: retrieves a canned joke.
3. Judge: "What did you say earlier about your favorite book?" Machine: must *remember* past conversation = knowledge representation.

Lesson. Passing requires more than tricks. The test stresses memory, reasoning, and flexibility.

Concluding Remarks. The essence of AI is not copying humans exactly, but designing systems that can act *rationally*, doing the right thing in context.

Chapter 2: Intelligent Agents

Agents and Environments. Definition 2 (Agent). An agent perceives its environment through *sensors* and acts on it through *actuators*.

PEAS Framework. To specify an agent's task, define:

- Performance measure (how we score it).
- Environment (where it lives).
- Actuators (how it acts).
- Sensors (how it perceives).

Example: Automated Taxi

- Performance: safe, legal, fast trips.
- Environment: roads, traffic, weather.
- Actuators: steering, pedals, display, voice.
- Sensors: cameras, GPS, speedometer, microphone.

The Concept of Rationality. Definition 3 (Rational Agent). A rational agent selects the action that maximizes expected performance given its knowledge and percepts to date.

Note. Rational $\neq$ perfect. It simply means "best possible under the circumstances."

Calculative Example: The Vacuum World. Setup. Two rooms: A and B. Each can be clean or dirty. Agent can: move left, move right, suck dirt, or do nothing.

Performance Measure: +1 for each clean square per time step. **Sensors:** detect current room + whether dirty. **Actuators:** move left/right, suck.

Percept	Action
(In A, dirty)	Suck
(In A, clean)	Move Right
(In B, dirty)	Suck
(In B, clean)	Move Left

Walkthrough:

1. Start: $(A, dirty)$. Agent sucks $\rightarrow$ square A is clean.
2. Now: $(A, clean)$. Agent moves right $\rightarrow$ goes to square B.
3. If $(B, dirty)$, agent sucks $\rightarrow$ both clean. Success!

Lesson. This *simple reflex agent* works well in this small world. But if dirt reappears or rooms are unknown, it may loop forever. Stronger agents (model-based or learning) perform better.

Pitfall Example: When Reflex Agents Fail. Scenario. Suppose our vacuum agent has no memory of past states. Its only percept is "dirty" or "clean" in the current square.

Problem. After cleaning one square, if it finds the square clean, it must *guess* what to do next. If it always moves left, but it started in the leftmost square, it will bump into the wall forever. If it always moves right, the same problem occurs on the other side.

Outcome. The agent can get stuck in an *infinite loop*, achieving very low performance. Even though the design looked correct in the table, the absence of memory makes it fragile.

Lesson. Simple reflex agents can succeed in tiny, cleanly defined worlds. But as soon as the world is partially observable or unpredictable, we need stronger designs:

- Model-based agents (remember past states).
- Learning agents (improve through feedback).

This pitfall motivates moving beyond reflexes to more powerful architectures.

Types of Agents.

1. Simple reflex agents: act only on current percept.
2. Model-based: maintain memory of unseen world parts.
3. Goal-based: choose actions to reach goals.
4. Utility-based: weigh tradeoffs (speed vs. safety).
5. Learning agents: improve performance over time.

Concluding Remarks. Agent design steps:

- Use PEAS to specify the environment clearly.
- Define rationality using performance measures.
- Choose agent architecture (reflex, model-based, goal, utility, learning).

This gives us a rigorous but approachable way to formalize "intelligence."