Chapter 1: Introduction to AI

# Introduction to AI

Learning objectives:

* Definition of AI
* Foundational and Current Disciplines
* AI Ethics

## Definition of AI

There are many definitions of AI. Some are bad, some are good. However, the definition that I like most from John McCarthy

“The science and engineering of building intelligent machine”

The definition captures 2 ambitious goals:

* **Science goal**: understand what intelligence is

(accurate models of cognition are now the focus of cognitive science, neuroscience and psychology)

* **Engineering goal**: build/implement intelligent machine

(focus on developing methods that match or exceed human performance in certain domains, possibly by different means)

There are 4 perspectives laid out along 2 dimensions:

* **Vertical direction**: human vs rational (close to human vs close to optimal)
* **Horizontal direction**: thinking vs acting (thought/reasoning vs actions/behaviour)

|  |  |
| --- | --- |
| **Thinking Humanly**  Machines with minds, in the full and literal sense | **Thinking Rationally**  Studying the computations that enable perception, reasoning, and actions |
| **Acting Humanly**  Concerned with the automation of intelligent behaviour | **Acting Rationally**  Computers do things at which, for now, people are better |

**Rationale:** (Why we consider AI definitions by 2 dimensions)

* **Horizontal direction (thinking vs acting):** A machine could behave intelligently without thinking. Do we really need a machine have ability to think like us?
* **Vertical direction (human vs rational):** We exhibit intelligence, but our intelligence is not perfect. Our behaviour is not always optimal. Do we want the machine to behave like us or optimally?

### Acting Humanly: The Turing Test approach

**Turing Test**:

**Purpose:** designed to provide a satisfactory operational definition of intelligence   
(the measure of how satisfied with intelligence operation).

**Rule:** a computer passes the test if a human interrogator cannot tell whether the written responses come from a person or from a computer after posing some written questions.

**Focused on AI components:**

* Natural Language Processing
* Knowledge Representation
* Automated Reasoning
* Machine Learning

**Note**: Turing’s test deliberately avoided direct physical interaction between the interrogator and the computer because *the* ***physical simulation of a person*** is unnecessary for intelligence.

However, the so-called **total Turing Test** includes a video signal (physical interaction), so that the interrogator can test the perceptual abilities of subject (computer/robot) as well as the opportunity for the interrogator to pass physical objects “through the hatch”. To pass the **total Turing Test**, the computer will need:

* **Computer vision**: ability to perceive objects.
* **Robot**: ability to manipulate objects and move about.

**Note**: It is important to study the underlying principles of intelligence (thinking humanly) than to duplicate an exemplar (in this case, mimic the human). Ex: Human tried to create plane by mimic bird’s wing motion. But it failed, the plane is created only when human understand why the bird can flight (aerodynamics).

### Thinking humanly: The cognitive modelling approach

**Purpose/Objective:** develop scientific theories of the activities of the brain

Three ways to get inside the actual workings of human minds:

* **Introspection:** trying to catch our own thoughts as they go by
* **Psychological experiments:** observing person in action
* **Brain imaging:** observing brain in action

Once we have sufficiently precise theory of the mind, it becomes possible to express the theory as a computer program (Artificial Intelligence Machine).

((program’s input, program’s output) == people behaviour) ⇔ some of the program’s mechanisms could also be operating in humans

Example: Allen Newell and Herbert Simon, who developed GPS, the “General Problem Solver”, were not content to have the program solve problems correctly but more concerned with comparing the trace of its reasoning steps to traces of human subjects solving the same problems.

Two approach to modelling cognition:

* **Cognitive Science (top-down):**

Uses computer models and experimental psychology techniques to predict and test behaviour of human subjects.

* **Cognitive Neuroscience (bottom-up):**

Uses computer imaging & other neurological data to observe brain in action.

These days, both disciplines are distinct from AI. (These disciplines are used with computer science to model cognition of human)

Note:

**Cognitive science** is a branch of psychology and is study of cognition or thought. It includes language, problem-solving, decision-making, and perception, especially consciously aware perception. Cognitive science started with those higher-level behavioural traits that were observable or testable and asked what is going on inside the mind or brain to make that possible.

**Neuroscience**is a branch of biology that began as the study of the anatomy and physiology of neural tissue. It grew out of clinical neurology and neurobiology, which evolved into neuroscience. Neurobiology concerns itself first and foremost with the observed anatomy and physiology of the brain, from major structures down to neurons and molecules. Neuroscience adds to that the study of how the brain works, mechanistically, functionally, and systemically to produce observable behaviour.

### Thinking Rationally: Laws of Thought

**Purpose/Objective:** formalise and mechanise valid reasoning (maths and philosophy 🡪 AI)

* **Formalisation step:** develop a precise notation for statements about all kinds of objects in the world and the relations among them. (≠arithmetic notation which only provides for statement about numbers)
* **Mechanisation step:** build programs that could solve any solvable problem described in logical notation.

**Logic:** notation and rules to derive valid conclusions:

* Aristotle’s syllogism.
* Mathematical development of classical logic. (arithmetic notation)
* Non-classical logic to formalise. (the notation to describe the world). Example: Sparrow is a bird, Birds fly 🡪 Sparrow fly

**Limitation of approach:**

* Limit 1: Undecidability

Goedl’s theorem: every axiomatizable consistent theory extending arithmetic has formulas that are true but not provable within the theory.

* Limit 2: Complexity
  + Not easy to take informal knowledge and state it in the formal terms required by logical notation (particular, when the knowledge is less than 100%)
  + Big difference between solving a problem in principle and solving in practice (problem with few hundred facts would exhaust the computational resources of any computer unless it has some guidance as to which reasoning steps to try first)
* Limit 3: Scope

Not all intelligent behaviour requires reasoning

* Limit 4: Purpose

Why do you need reasoning?

### Acting rationally: The rational agent approach

**Computer programs** do something

**Computer agents** do more:

* Operate autonomously
* Perceive their environment
* Persist over a prolonged time period
* Adapt to change
* Create and pursue the goal

**Rationality:** is about doing the right thing

**A rational agent:** acts so as to achieve the best outcome or when there is uncertainty, the best expected outcome, given the information available and time available (limited rationality)

Note: Making correct inferences (from laws of thought approach) sometimes is a part of agent because one way to act rationally is to reason logically to the conclusion that a given action will achieve one’s goals and then to act on that conclusion. However, correct inference is not all of rationality, there is no provably correct thing to do, but something must still be done. There are also the ways of acting rationally that cannot be said to involve inference. For example: recoiling from a hot stove is a reflex rational action without any inference. Besides, all skills needed for Turing test also allow agent to act rationally (Knowledge representation and reasoning enable agents to reach good decisions).

**Advantage:**

* More general than the laws of thought approach because correct inference is just one of several possible mechanisms for achieving rationality.
* More amenable to scientific development than are approaches based on human behaviour or human thought.

## Foundational and current disciplines

I take these slides from ANU slide COMP6320

Graphical user interface, text

Description automatically generated

Text

Description automatically generated with medium confidence

## AI ethics and risks

**AI risk:**

* People might lose their jobs

However, AI create wealth, does dangerous and boring jobs for us. But with the development of automation, due to the high accuracy and lower cost, many companies would switch to AI instead of hiring employees.

* Accountability loss

If a doctor depends on AI medicine expert to make a wrong diagnosis who will take the responsibility – the people who create AI expert or doctor.

* AI reproducing negative biases and attitudes

There are many sources of negative bias information that would be used to create AI agents. These AI agents will be then applied widely over the world and unfortunately make the negative bias attitude spreading. AI should only share our positive value

* Use of AI as weapon

Misuse of AI in many circumstances could make catastrophes. Such as: AI could be used in cyberattack, intelligence robot used in the war

* AI Success might end of human era

Once machine surpasses the human, it can create smarter machines.

Intelligence explosion and singularity at which human era ends. (singularity is an explosion point of AI, when AI reach that point, its intelligence will increase dramatically)

However, there are many **counters** such as:

* Limit to intelligence
* Nothing special about human intelligence
* Computational complexity
* “intelligence to do task” ≠ “the ability to improve the intelligence to do task”

**AI ethics**

We have robotics laws: (this slide is taken from COMP6320 - ANU course)

A picture containing text

Description automatically generated