
CM0312 Artificial Intelligence

I Intelligent Systems

I.1 Machines with Minds

F. C. Langbein

School of Computer Science
and Informatics
Cardiff University



Version 1.4

Overview

- Machines with minds: What is Artificial Intelligence?
 - Acting humanly: Turing test, Searle's Chinese room
 - Acting rationally: Aristotle's algorithm
 - Thinking rationally: logic
 - Thinking humanly: cognitive science, neuroscience
- Our goal: intelligent agents
 - Model: agent, environment, task
 - Control: agent function
 - Vacuum cleaner example

F. C. Langbein, CM0312 Artificial Intelligence – I Intelligent Systems; I.1 Machines with Minds

1

Can Machines have a Mind?!

"The fundamental goal of this research is not merely to mimic intelligence or produce some clever fake. Not at all. AI wants only the genuine article: machines with minds, in the full and literal sense. . .

Scoffers find the whole idea quite preposterous — not just false but ridiculous — like imagining that your car (really) hates you or insisting that a murderous bullet should go to jail.

Boosters. . . are equally certain that it's only a matter of time; computers with minds, they say, are as inevitable as interplanetary travel and two-way pocket TV."

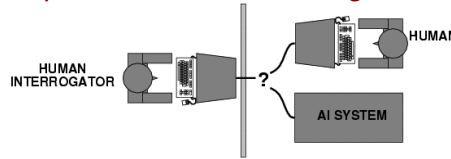
J. Haugeland, 1985

F. C. Langbein, CM0312 Artificial Intelligence – I Intelligent Systems; I.1 Machines with Minds

2

Turing Test

- Turing (1950): “Can machines think?”
 - ➔ “Can machines behave intelligently?”
- Turing Test: *operational test for intelligent behaviour*



- ➔ Suggests all major AI components:
 - **knowledge**, **reasoning**, **learning**
 - also language understanding, vision, robotics, . . .
- Is a machine that passes this test really intelligent?
 - Aim is to **understand principles**, not simulate exemplar

Searle's Chinese Room

- **System**: Human with book and paper inside room with small slot
 - **CPU**: Human who understands English
 - **Program**: Rule book written in English
 - **Memory**: Stacks of blank paper, some with *indecipherable* symbols
- **Process**: Paper with indecipherable symbols inserted through slot
 - Human finds symbols in rule book and follows instructions:
 - write new symbols on paper
 - find symbols in the stacks
 - rearrange stacks, etc.
 - Leads to symbols on a piece of paper output through the slot
- This process produces reasonable answers to questions in Chinese
 - So it passes the Turing test, but is it intelligent?
 - According to Searle **no understanding** is taking place

Ethics or an Early Algorithm

- (A)I **was** is also a problem for humans. . .

“We deliberate not about ends, but about means. For a doctor does not deliberate whether he shall heal, nor an operator whether he shall persuade, . . . They assume the end and consider how and by what means it is attained, and if it seems easily and best produced thereby; while if it is achieved by one means only they consider how it will be achieved by this and by what means this will be achieved, till they come to the first cause, which in the order of discovery is last. . . and what is last in the order of analysis seems to be first in the order of becoming. And if we come on an impossibility, we give up the search, e.g. if we need money and this cannot be got. . . ”

Aristotle, Nicomachean Ethics, 328BC

- ➔ Implemented 2300 years later by Newell and Simon's GPS

What is the Right Thing to Do?

- Rational behaviour: **Doing the right thing**
 - The “right thing”:
 - that which is expected to maximise goal achievement, given the available information**
 - Thinking not necessarily required (e.g. blinking reflex)
 - Thinking should be in the service of rational action
- Aristotle (Nicomachean Ethics):
“Every art and every inquiry, and similarly every action and pursuit, is thought to aim at some good”
- Devise systems *acting in a rational manner towards a goal*
 - Try to measure goal achievement
- Caution: Humans do not act like this!

Thinking Rationally

- Aristotle: “what are **correct arguments/thought processes**?”
 - Normative/prescriptive rather than descriptive
 - What is the purpose of thinking?
 - What thoughts *should* I have?
 - Greek schools developed various forms of *logic*:
 - notation and rules of derivation for thoughts**
 - Direct line through mathematics and philosophy to AI
- Problems:
 - (1) *Informal, uncertain knowledge* hard to formalise in logic
 - (2) *Computationally expensive*: while programs can solve problems *in principle*, it may take a very long time
- Science explains experiments, and does not prescribe

How do Humans Think?

- Study the brain as information processing device
- Scientific theories of **internal activities of the brain**:
 - Different levels of abstraction:
 - *Knowledge, behaviour, environment*
 - Cognitive science, psychology (top-down)
 - Predicting and testing of human behaviour
 - *Neural networks, synapses, biochemical reaction networks, DNA, hormones, ...*
 - Neuroscience, biology, chemistry (bottom-up)
 - Identification from neurological, etc. data
- As in AI, no theory explains anything resembling human-level general intelligence

What is Artificial Intelligence?

- Approaches to Artificial Intelligence:
Turing test, rational behaviour, logic, science of the brain
- Systems that . . .

act like humans	act rationally
"The art of creating machines that perform functions that require intelligence when performed by people." [Kurzweil, 1990]	"AI... is concerned with intelligent behaviour in artifacts." [Nilsson, 1998]
think like humans	think rationally
"[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning..." [Bellman, 1978]	"The study of computations that make it possible to perceive, reason, and act." [Winston, 1993]

Weak AI

Strong AI

Our Goals

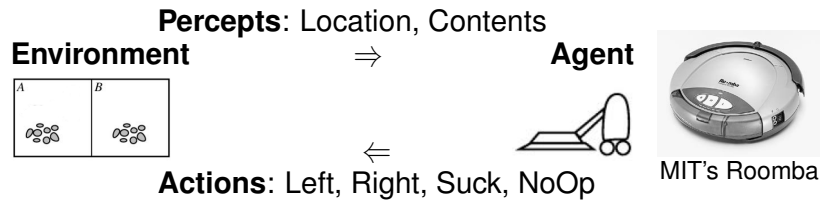
- **Scientific goal:** to understand the principles that make intelligent behaviour possible, in natural or artificial systems
- **Engineering goal:** to specify methods for the design of useful, intelligent artifacts
- **Understand the principles by studying artefacts** designed to be intelligent:
 - For given class of environment, task, machine,
 - we seek algorithm(s) with the best performance
- Caveat:
 - **Computational limits** make perfect rationality infeasible
 - Design best algorithm for available resources

Intelligent Agents

- The systems we study are modelled as **agents**:
anything which **perceives** and **acts** in an **environment**
 - Humans, robots, softbots, thermostats, . . .
- An **intelligent agent** acts intelligently:
 - acts appropriately for goals and circumstances
 - flexible to changing environments and goals
 - learns from experience
 - considers perceptual and computational limitations
- Actions are controlled by **agent function**

$$\mathcal{A}: P^* \rightarrow A, p^* \mapsto \mathcal{A}(p^*) = a$$
 - maps percept history $p^* \in P^*$ to action $a \in A$

Vacuum Cleaner Model



- **Goals:** clean carpet, do not cause damage, do not break down, ...
- **Prior knowledge:** categories of objects (what is dirt?), what a sensor tells, ...
- **Past experiences:** effects of moving, cleaned room in the past, what happens when certain objects are sucked in, ...

Vacuum Cleaner Control

- **Agent function** maps from percept histories to actions:
 $\mathcal{A}: P^* \rightarrow A$

Percept history	Action	
(A, Clean)	Right	➤ What is the best function?
(A, Dirty)	Suck	
(B, Clean)	Left	➤ Can it be computed by an algorithm?
(B, Dirty)	Suck	
(A, Clean), (A, Clean)	Right	➤ Can it be executed within resource limits?
(A, Clean), (A, Dirty)	Suck	
:		

```
function action ← VACUUM-CLEANER (location, status)
  if status = Dirty, return Suck
  else if location = A, return Right
  else if location = B, return Left
```