Artificial Intelligence I

Dr Mateja Jamnik

Computer Laboratory, Room FC18

Telephone extension 63587

Email: mj201@cl.cam.ac.uk

http://www.cl.cam.ac.uk/users/mj201/

Notes I: General introduction to artificial intelligence and agents

Introduction: what's AI for?

Homo Sapiens = "Man the wise"

What is the purpose of Artificial Intelligence (AI)?

- To *understand intelligence* and to *understand ourselves*. This aim is shared by philosophy and psychology.
- To *make* intelligent systems. More exclusively the realm of CS.
- To make and sell cool stuff!

Our brain is small and slow, so why is it so good? (Actually this claim is of highly dubious accuracy, although it's often repeated.)

Introduction: what's the character of the field?

In many ways this is a young field (1956).

- This means we can actually *do* things!
- Also, we know what we're trying to do is *possible*.
- On the down side, it tends to mean that any perceived lack of success tends to be given more weight than is appropriate.

Philosophy has addressed such problems for at least 2000 years.

- *Can* we do AI? Is AI even possible at all?
- *Should* we do AI?

Arguably, philosophy has had relatively little success. Perhaps the most important open problem the world has left?

Introduction: what's happened since 1956?

Computers have taken us from theory to practice.

The simple ability to *try things out* has led to huge advances in a relatively short time.

- Perception (vision, speech processing...)
- Logical reasoning (prolog, expert systems...)
- Playing games (chess, backgammon, go...)
- Diagnosis of illness (in various contexts...)
- Theorem proving (assorted mathematical results...)
- Literature and music (automated writing and composition...)
- Robotics (a wide assortment of devices and applications...)

Introduction: what is the nature of the pursuit?

What is AI?

Well, it depends on who you ask...

We can find many definitions and a rough categorisation can be made depending on whether we are interested in:

- the way in which a system *acts* or the way in which it *thinks*, and;
- whether we want it to do this in a *human* way or a *rational* way.

Here, the word *rational* has a special meaning: it means doing the *correct* thing in given circumstances.

Acting like a human

Alan Turing proposed what is now known as the *Turing Test*.

- A human judge is allowed to interact with an AI program via a terminal.
- This is the *only* method of interaction.
- If the judge can't decide whether the interaction is produced by a machine or another human then the program passes the test.

In the *unrestricted* Turing test the AI program may also have a camera attached, so that objects can be shown to it, and so on.

Acting like a human

The Turing test is informative, and (very!) hard to pass.

- It requires many abilities that seem necessary for AI, such as learning. BUT: a human child would probably not pass the test!
- Sometimes an AI system needs human-like acting abilities—for example *expert systems* often have to produce explanations—but *not always*.

See the Loebner Prize:

`http://www.loebner.net/Prizef/loebner-prize.html`

Aside I: computer engineering (1940 to present)

To have AI, you need a means of *implementing* the intelligence. Computers are (at present) the only devices in the race! (Although quantum computation is looking interesting...)

AI has had a major effect on computer science:

- time sharing
- interactive interpreters
- linked lists
- storage management
- some fundamental ideas in object-oriented programming
- an so on...

When AI has a success, the ideas in question tend to stop being called AI!

Thinking like a human

There is always the possibility that a machine *acting* like a human does not actually *think*.

The *cognitive modelling* approach to AI has tried to:

- deduce *how humans think*—for example by *introspection* or *psychological experiments*—and,
- copy the process by mimicking it within a program.

An early example of this approach is the *General Problem Solver* produced by Newell and Simon in 1961. They were concerned with whether or not the program reasoned in the same manner that a human did.

$$\text{Computer Science} + \text{Psychology} = \textit{Cognitive Science}$$

Aside II: philosophy (428 B.C. to present) and psychology (1879 to present

Socrates wanted to know whether there was an algorithm (!) for "piety", prompting Plato to consider the rules governing rational thought.

This led to the *syllogisms*.

The possibility of reasoning being done *mechanically*: Ramon Lull's *concept wheels* (approx. 1315).

Various other attempts at mechanical calculators.

Aside II: philosophy

Mind as a *physical system*: Rene Descartes (1596-1650).

- is *mind* distinct from *matter*?
- what is *free will*?

Dualism: part of our mind—-the *soul* or *spirit*— is set apart from the rest of nature.

Aside II: philosophy

The opposing position of *materialism* was taken up by Wilhelm Leibnitz (1646-1716).

He attempted to build a machine to perform mental operations but failed as his logic was too weak.

(There is an intermediate position: mind is physical but unknowable.)

Aside II: philosophy

If mind is physical where does *knowledge* come from?

Francis Bacon (1561-1626): *empiricism*.

↓

John Locke (1632-1704): “Nothing is in the understanding, which was not first in the senses”.

In *A Treatise of Human Nature*, David Hume (1711-1776) introduced the concept of *induction*: we obtain rules by repeated exposure.

This was developed by Bertrand Russel (1872-1970): *observation sentences* are connected to *sensory inputs*, and all knowledge is characterised by logical theories connected to these. *Logical positivism*.

The *nature* of the connection between theories and sentences is the subject of Rudolf Carnap and Carl Hempel’s *confirmation theory*.

Aside II: philosophy

Finally: what is the connection between *knowledge* and *action*? How are actions *justified*?

Aristotle: don't concentrate on the *end* but the *means*.

If to achieve the end you need to achieve something intermediate, consider how to achieve that, and so on.

This approach was implemented in Newell and Simon's 1972 GPS.

Aside II: psychology

Modern psychology (arguably) began with the study of the human visual system performed by Hermann von Helmholtz (1821-1894).

The first *experimental psychology* lab was founded by his student Wilhelm Wundt (1832-1920) at the University of Leipzig.

- The lab conducted careful, controlled experiments on human subjects.
- The idea was for the subject to perform some task and *introspect* about their thought processes.

Other labs followed this lead. BUT: a strange—and fatal—effect appeared.

For each lab, the introspections of the subjects turned out to conform to the preferred theories of the lab!

Aside II: psychology

The main response to this effect was *behaviourism*, founded by John Watson (1878-1958) and Edward Lee Thorndike (1874-1949).

- They regarded evidence based on introspection as fundamentally unreliable, so...
- ...they simply rejected all theories based on any form of mental process.
- They considered only *objective* measures of *stimulus* and *response*.

Learnt a LOT of interesting things about rats and pigeons!

Aside II: psychology

The (arguably somewhat more sophisticated) view of the brain as an *information processing device*—the view of cognitive psychology—was steamrollered by behaviourism until Kenneth Craik's *The Nature of Explanation* (1943).

The idea that concepts such as reasoning, beliefs, goals *etc* are important is re-stated.

Critically: the system contains a model of the world and of the way its actions affect the world.

Aside II: psychology

stimuli converted to internal representation

↓

cognitive processes manipulate internal representations

↓

internal representations converted into actions

Thinking rationally: the “laws of thought”

The idea that intelligence reduces to *rational thinking* is a very old one. Aristotle first tried to model thought this way through *syllogisms*.

The general field of *logic* made major progress in the 19th and 20th centuries, allowing it to be applied to AI.

- we can *represent* and *reason* about many different things;
- The *logicist* approach to AI.

Aside III: mathematics (800 to present)

Philosophers have had some great ideas, but to be *scientific* about AI three areas of mathematics are needed: computation, logic, and probability.

Logic:

- To the likes of Aristotle, a philosophical rather than mathematical pursuit.
- George Boole (1815-1864) made it into mathematics.
- Gottlob Frege (1848-1925) founded all the essential parts of first-order logic.
- Alfred Tarski (1902-1983) founded the theory of reference: what is the relationship between *real* objects and those in logic.

Aside III: mathematics

Computation:

- Concept of an algorithm: Arab mathematician *al-Khowarazmi*. On Calculation with Hindi Numerals, 825 AD.
- What are the limits of algorithms? David Hilbert's (1862-1943) *entscheidungsproblem*.
- Solved by Turing, who (with others) formulated precisely what an algorithm *is*.
- Ultimately, this has lead to the idea of *intractability*.
- Kurt Godel (1906-1978): theorems on completeness and incompleteness.

Thinking rationally: the “laws of thought”

Unfortunately there are obstacles to any naive application of logic. It is hard to:

- represent *commonsense knowledge*;
- deal with *uncertainty*;
- reason without being tripped up by *computational complexity*.
- sometimes it’s necessary to act when there’s no logical course of action;
- sometimes inference is unnecessary (reflex actions).

Aside IV: probability (1501 to present)

Probability:

- Gerolamo Cardano (1501-1576): gambling outcomes.
- Further developed by Fermat, Pascal, Bernoulli, Laplace...
- Bernoulli (1654-1705) in particular proposed probability as a measure of *degree of belief*.
- Bayes (1702-1761) showed how to update a degree of belief when new evidence is available.
- Probability forms the basis for the modern treatment of uncertainty.
- The decision theory of Von Neumann and Morgenstern (1944) combines uncertainty with action.

Acting rationally

Basing AI on the idea of *acting rationally* means attempting to design systems that act to *achieve their goals* given their *beliefs*.

What might be needed?

- To make *good decisions* in many *different situations* we need to *represent* and *reason* with *knowledge*.
- We need to deal with *natural language*.
- We need to be able to *plan*.
- We need *vision*.
- We need *learning*.
- We need to *deal with uncertainty*.

This looks like a summary of modern AI!

Acting rationally

The idea of acting rationally has several advantages:

- the concepts of *action*, *goal* and *belief* can be defined precisely making the field suitable for scientific study, whereas;
- dealing with humans involves a system that is still changing and adapted to a very specific environment.

Also, all of the things needed to pass a Turing test seem necessary for rational acting.

Other contributions I: linguistics (1957 to present)

B. F. Skinner's *Verbal Behaviour* (1951) set out the approach to *language* developed by the behaviourists.

It was reviewed by Noam Chomsky, author of *Syntactic Structures*:

- He showed that the behaviourists could not explain how we understand or produce sentences that we have *not previously heard*.
- Chomsky's own theory—based on syntactic models as old as the Indian linguist Panini (350 B.C.), did not suffer in this way.
- Chomsky's own theory was also formal, and could be programmed.

Other contributions I: linguistics

This overall problem is considerably harder than was realised in 1957.

It requires knowledge representation, and the fields have informed one another.

“Time flies like an arrow”

“Fruit flies like a banana”

Other contributions II: economics (1776 to present)

How should I act, perhaps in the presence of adversaries, to obtain something nice in the future?

- Adam Smith: *An Inquiry into the Nature and Causes of the Wealth of Nations* (1776).
- When we say “something nice,” how can the “degree of niceness” be measured?

  This leads to the idea of *utility* as a mathematical concept.

  Developed by Leon Walras (1834-1910), Frank Ramsey (1931) and John Von Neumann and Oskar Morgenstern (1944).

Other contributions II: economics

- For *large* economies:

  Probability theory + utility theory = decision theory

- *Game theory* is more applicable to *small* economies.

  In some games it turns out to be *rational* to act (apparently) randomly.

- Dealing with *future* gains resulting from a sequence of actions: operations research and *Markov decision processes*, the latter due to Richard Bellman (1957).

Unfortunately it is computationally hard to act rationally.

Herbert Simon (1916-2001) won the Nobel Prize for Economics in 1978 for his work demonstrating that *satisficing* is a better way of describing the actual behaviour of humans.

Other contributions III: neuroscience (1861 to present)

Nasty bumps on the head

↓

We know that the brain has something to do with consciousness

Experiments by Paul Broca (1824-1880) led to the understanding that localised regions have different tasks.

Around that time the presence of *neurons* was understood *but* there were still major problems.

For example, even now there is no complete understanding of how our brains store a single memory!

More recently: EEG, MRI and the study of single cells.

Other contributions IV: cybernetics and control theory (1948 to present)

Ktesibios of Alexandria (250 BC)

The first machine to be able to modify its own behaviour was a water clock containing a mechanism for controlling the flow of water.

- James Watt (1736-1819): governor for steam engines
- Cornelius Drebbel (1572-1633): thermostat
- Control theory as a mathematical subject: Norbert Wiener (1894-1964) and others.

This presented another challenge to behaviourism.

Other contributions IV: cybernetics and control theory

Interesting behaviour caused by a *control system* minimising *error*

error = difference between *goal* and *current situation*

More recently, we have seen *stochastic optimal control* dealing with the maximisation over time of an *objective function*.

This is connected directly to AI, but the latter moves away from *linear*, *continuous* scenarios.

What's in this course?

This course introduces some of the fundamental areas that make up modern AI:

- An outline of the background to the subject.
- An introduction to the idea of an agent.
- Solving problems in an intelligent way by search.
- Playing games.
- Knowledge representation, and reasoning.
- Learning.
- Planning.

If time, a little philosophy. (A crash course on how to survive at parties!)

What's *not* in this course?

- Nothing is said about the classical AI programming languages Prolog and Lisp.
- A great deal of all the areas on the last slide!
- Perception: vision, hearing and speech processing, touch (force sensing, knowing where your limbs are, knowing when something is bad), taste, smell.
- Natural language processing
- Acting on and in the world: robotics (effectors, locomotion, manipulation), control engineering, mechanical engineering, navigation.
- Genetic algorithms.
- Fuzzy logic.
- Uncertainty and much further probabilistic material.

Text books and prerequisites

The course is based on the relevant parts of:

Artificial Intelligence: A Modern Approach, Second Edition (2003). Stuart Russell and Peter Norvig, Prentice Hall International Editions.

The prerequisites for the course are:

- A little logic.
- Algorithms and data structures.
- Discrete and continuous mathematics.
- Basic computational complexity.

Interesting things on the web

- Winning the DARPA Grand Challenge with an AI Robot:

  `ai.stanford.edu/~dstavens/aaai06/montemerlo_etal_aaai06.pdf`

- General resource page for machine learning:

  `home.earthlink.net/~dwaha/research/machine-learning.html`

- The Cyc project:

  `www.cyc.com`

- Human-like robots:

  `www.ai.mit.edu/projects/humanoid-robotics-group/`

- Sony robots: !!!!!DISCONTINUED

  `www.aibo.com`

- Honda “ASIMO”:

  `world.honda.com/ASIMO`

- NEC “PaPeRo”:

  `www.incx.nec.co.jp/robot`

Agents

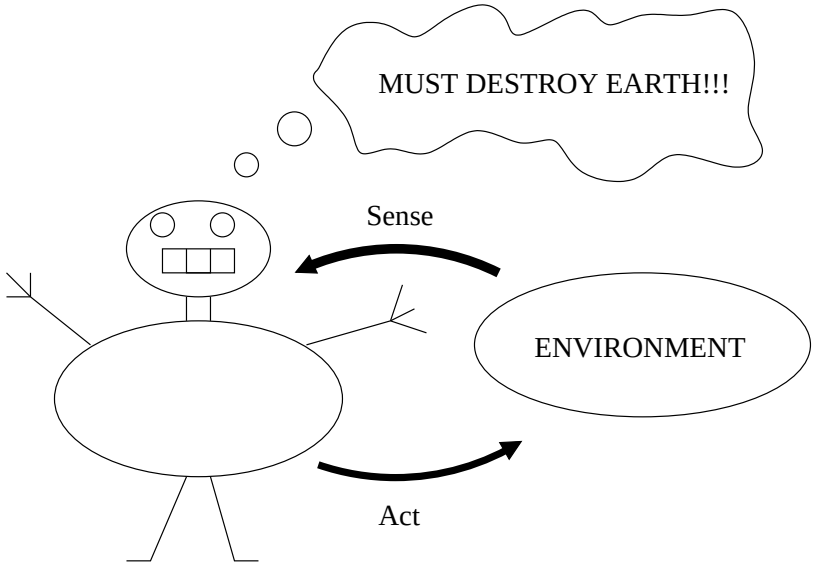
We now look at a simple unifying concept for the construction of AI systems: the idea of an *agent*. **Aims:**

- to introduce agents as a way of speaking about a wide range of AI systems;
- to look at some ways in which agents might be structured;
- to connect these structures to standard fields of study within AI as a subject;
- to look briefly at ways in which an agent's *environment* is significant.

Reading: Russell and Norvig, chapter 2.

Agents

There are many different definitions for the term *agent* within AI. (Be aware of this when reading beyond the course textbook.)



We will use the following simple definition: *an agent is any device that can sense and act upon its environment*.

Agents

This definition can be very widely applied: to humans, robots, pieces of software, and so on. It is only one of many.

Questions:

- How can we judge an agent's performance?
- How can we begin successfully to design an agent?
- How can an agent's environment affect its design?

Recall that we are interested in devices that *act rationally*, where 'rational' means doing the *correct thing* under *given circumstances*.

Measuring performance

Clearly any *performance measure* we apply will need to be domain-dependent, so we will have to design one appropriate for a given problem.

Example: for a chess playing agent, we might use its rating.

Example: for a mail-filtering agent, we might devise a measure of how well it blocks spam, but allows interesting email to be read.

Example: for a car cleaning robot, we might want maximum removal of dirt in minimum time. Being more sophisticated, perhaps we don't want the car to get damaged, or to use too much water or energy, and we might want the robot to have spare time at the weekend to write its novel...

So: the choice of a performance measure can be tricky.

Measuring performance

Two further points:

- In general, we will be interested in success *over the long term*. For example, we might not want to favour a car-cleaner that's extremely fast in the first hour and then sits around reading, over one that works consistently.
- We are generally interested in *expected* performance because usually agents are not *omniscient*—they don't *infallibly* know the outcome of their actions.

It is *rational* for you to enter this lecture theatre even if the roof falls in today.

Measuring performance

So rational behaviour requires us to know:

- A well-defined measure of performance.
- What our agent has already perceived. The *percept sequence*.
- What our agent knows about the environment it lives in.
- What actions our agent is capable of performing.

An *ideal rational agent* acts as follows: for any percept sequence it acts so as to expect to maximise performance, given what it knows about the world via the percept sequence and its own knowledge.

So: an agent capable of detecting and protecting itself from a falling roof might be more *successful* than you, but *not* more *rational*.

Environments

Some common attributes of an environment have a considerable influence on agent design.

- **Accessible/inaccessible:** do percepts tell you *everything* you need to know about the world?
- **Deterministic/non-deterministic:** does the future depend predictably on the present and your actions?
- **Episodic/non-episodic** is the agent run in independent episodes.
- **Static/dynamic:** can the world change while the agent is deciding what to do?
- **Discrete/continuous:** an environment is discrete if the sets of allowable percepts and actions are finite.
- **Single-agent/multi-agent:** is the agent acting individually or in the presence of other agents. In the latter case is the situation *competitive* or *cooperative*, and is *communication* required?

Basic structures for intelligent agents

Example: email spam filter.

Percepts: the textual content of individual email messages. (A more sophisticated program might also take images or other attachments as percepts.)

Actions: send to the inbox, delete, or ask for advice.

Goals: remove spam while allowing valid email to be read.

Environment: an email program.

Basic structures for intelligent agents

Example: aircraft pilot.

Percepts: sensor information regarding height, speed, engines *etc*, audio and video inputs, and so on.

Actions: manipulation of the aircraft's controls. Also, perhaps talking to the passengers *etc*.

Goals: get to the current destination as quickly as possible with minimal use of fuel, without crashing *etc*.

Environment: aircraft cabin.

Programming agents

A basic agent program is as follows:

```
action agent(percept)
{
   static memory;      // the agent's memory.

   memory = update_memory(memory,percept);
   next_action = choose_action(memory);
   memory = update_memory(memory,next_action);

   return next_action;
}
```

It is up to the agent how long a list of past percepts it stores, and the measure of performance is not known to the agent.

Programming agents

The simplest approach would be to use a table to map percept sequences to actions, or we could produce a program capable of reproducing such a table, but this can quickly be rejected.

- The table will be *huge* for any problem of interest. About 35^{100} entries for a chess player.
- We don't usually know how to fill the table.
- Even if we allow table entries to be *learned* it will take too long.
- The system would have no *autonomy*.

We can overcome these problems by allowing agents to *reason*.

Autonomy

If an agent's behaviour depends in some manner on its *own experience of the world* via its percept sequence, we say it is *autonomous*.

- An agent using only built-in knowledge would seem not to be successful at AI in any meaningful sense: its behaviour is predefined by its designer.
- On the other hand *some* built-in knowledge seems essential, even to humans.

Not all actual animals are entirely autonomous. **For example:** dung beetles.

Reflex agents

We can’t base our example spam filter on a table: there are too many character sequences to consider.

But we might try *extracting pertinent information* and using *rules* based on this.

Condition-action rules:

if a certain *state* is observed **then** perform some action

Example:

if message contains ‘gambling’ and ‘online’ **then** delete

Keeping track of the environment

Some points immediately present themselves regarding reflex agents:

- we can't always decide what to do based on the current percept;
- however storing *all* past percepts might be undesirable (for example requiring too much memory) or just unnecessary;
- reflex agents don't maintain a description of the state of their environment;
- however this seems necessary for any meaningful AI. (Consider automating the task of driving.)

This is all the more important as usually percepts don't tell you everything about the state.

Keeping track of the environment

An agent should maintain:

- a description of the current state of its environment;
- knowledge of how the environment changes independently of the agent;
- knowledge of how the agent's actions affect its environment.

This requires us to do *knowledge representation* and *reasoning*.

Goal-based agents

Sometimes, choosing a rational course of action depends on your *goal*.

- We need to consider *goal-based agents*.
- As agents include knowledge of how their actions affect the environment, they have a basis for choosing actions to achieve goals.
- To obtain a *sequence* of actions we need to be able to *search* and to *plan*.

This is *fundamentally different* from a reflex agent. For example, by changing the goal you can change the entire behaviour.

Utility-based agents

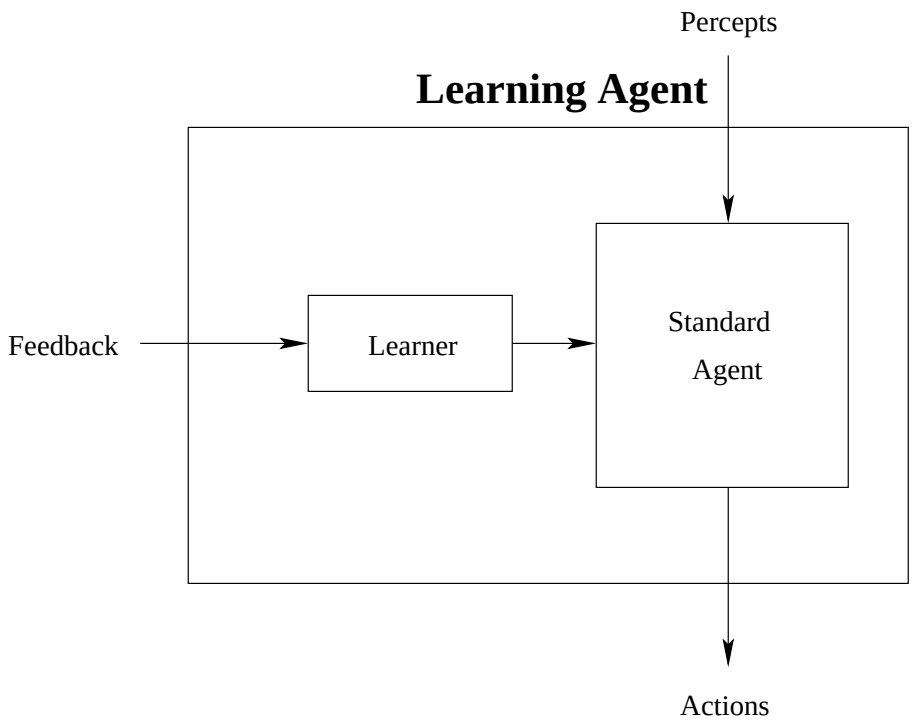
Introducing goals is still not the end of the story.

There may be many sequences of actions that lead to a given goal, and some may be preferable to others.

A *utility function* maps a state to a number representing the desirability of that state.

- We can trade-off *conflicting goals*, for example speed and safety.
- If an agent has several goals and is not certain of achieving any of them, then it can trade-off likelihood of reaching a goal against the desirability of getting there.

Learning agents



Here, the *learner* needs some form of *feedback* on the agent's performance. This can come in several different forms. In general, we also need a means of *generating new behaviour* in order to find out about the world.