Today

More philosophical issues about AI

- behaviourism
- strong and weak versions of AI
- Searle's Chinese room
- Penrose and non-computability

Free will ctd

The view that free-will is simply an illusion as normally understood is simply an illusion has been expressed many times, eg:

> Men believe themselves to be free, because they are conscious of their own actions and are ignorant of the causes by which they are determined.
> Spinoza, Ethics, book 3

and so

> The murderer is no more responsible for his or her behaviour than is a river that floods a village (Spinoza)

More modern attitudes

We should remember that terms like "intelligence", "free will" are not clearly defined, and indeed neither is it clearly defined what it is to be a person (to be you, me, him, her).

Building different agant systems give an idea of what different agent architectures support in terms of eg deliberative vs reactive agents (see Sloman's work of architecture of mind).

A good discussion (from a particular viewpoint) is in

> Daniel Dennet
> Elbow Room: the Varieties of Free Will Worth Wanting
> OUP 1984

Behaviourist psychology

Suppose we want to build an account of the mental functioning of people. We have two sorts of information to work with:

- What we can observe of their behaviour, in various environments.
- What we know about our own thoughts, feelings, desires, motivations.

Can we build a psychological theory without talking about the second class of information (internal mental states of believing, desiring, thinking, *etc*)? Behaviourist psychology tried to do this.

Objects with desires

In some cultures, it was/is common to suppose that all sorts of objects had desires. For example "A stone falls to the ground when dropped because it is an earth-like objects and it seeks its natural place."

Progress is physics came about when an explanation became available that talked only about observable behaviour.

In the same way, it was thought that a psychological theory could be built just taking into account what can be observed from the outside.

Why follow behaviourism?

At first sight it seems strange to try to understand mental activity while ignoring what we know about our own thoughts. Why should we do this?

- We can get agreement on what external behaviour is, but we cannot be sure what the thoughts of other people are.
- We understand other people's minds by looking at their behaviour – we have no direct access to their minds.

So the behaviourist will give a meaning for a mental word (*eg* "pain") in terms of a set of behaviours (grimace, cry of 'ow!', . . .).

Behaviourist accounts of learning

Typical behaviourist accounts of human behaviour are in terms of stimulus and response, and in terms of reinforcement of behaviour by some reward for "correct" actions. Descriptions of the training of a neural net fit easily into this framework.

For people, it is hard to devise an account of how they think internally that can be shown to be right or wrong in general.

However, for an artificial system, we may be in a quite different situation, where we have designed the system ourselves, and have built it to work in terms of some internal states that we use to describe its actions.

Developments like this have led to new philosophical theories of mental functioning.

Strong and Weak AI

In the last lecture, we looked two ways of thinking about AI systems:

- looking just at behaviour;
- considering internal states.

Today we want to consider the relation between computer *simulations* of understanding, intelligence, etc., and (real!) artificial understanding, intelligence, etc.

Simulation

Computer modelling is used in many domains, to provide a way of predicting behaviour of some systems (physical, ecological, economic . . .).

For example, a computer program can be used to take meteorological information from the last 24 hours, and estimate the evolution of the weather for the next 24 hours. This uses the known principles about the evolution of weather patterns in time; the *simulation* consists in computing successively the weather configuration for every successive time separated by some fixed interval.

However, it's not generally suggested that this simulation constitutes *artificial weather*.

Describing Algorithms

An algorithm is a clearly described prescription for carrying out a computation, given some input.

One way of describing an algorithm is to make use of the idea of the *state* of the computing device. This is part of the general characterisation of computation due to Turing called the *Turing machine*. Amazingly, anything that can be computed on any digital computer can be computed via a Turing machine.

Simplified, this works as follows.

Turing Machine

The machine has a tape, which can be extended without bound. A single square is being scanned at any time.

We have an *alphabet* of characters that can be written on the tape, and a finite set of internal *states*, one of which is the *initial state*, and others are *stopping states*.

The machine has a set of instructions, each of the form

> If in state Q and looking at symbol S, replace S with S', and move one square to right (or left).

Many algorithms for one observable behaviour

Now suppose that what we can see of a computing device is limited to seeing the inputs, and the outputs (so we can't see the internal state). Usually there will be *many ways* to achieve this behaviour, using different states internally, or the same states in a different way.

How does this relate to the evolution of human mental states, in terms of our reaction to sensory input, and our mental state?

The philosopher Hilary Putnam suggested that psychological states are just computational states of the brain, conceived of as a computing device. (He has since changed his mind!)

Strong AI

We said earlier that we can't determine internal mental states just from external behaviour. However, it may be that we can completely describe human mental processes in terms of computing devices with (a huge set of) internal states, and a way of reacting to sensory input, dependent on internal state.

Notice that such a description says *nothing* about the physical device supporting this computation (eg human brain, silicon chip).

According to this view, mental processes (including understanding, consciousness) can in principle be described in state-based computational terms. Then *any* execution of such an algorithm *is* a conscious, understanding process.

Strong AI (ctd)

The term *strong AI* was introduced by the philosopher John Searle as follows:

> According to . . . this view, the brain is just a digital computer, and the mind is just a computer program. One could summarise this view – I call it strong AI – by saying that the mind is to the brain as the program is to computer hardware.
>
> Minds, Brains, and Science

Searle argues that this view is mistaken.

Weak AI

An alternative position is the following: it is possible to *simulate* human intelligence with a digital computer. That is

> . . . the view that brain processes (and mental processes) can be simulated computationally [is] Weak AI . . .
>
> Searle, The Rediscovery of the Mind

According to this view, computer models of human intelligence can only be like computer models of the weather; they give a way to predict the evolution of the system, but don't actually build intelligence, or weather.

Non-computability

Recently, a third position has been advocated by Roger Penrose; it may be that some of the physical processes in the brain cannot even be simulated by a digital computer.

> Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally.
>
> Penrose, Shadows of the Mind

We won't explore his reasons for believing this is the case, but this is a coherent position, distinct from both strong and weak AI positions.

Are we machines?

Although Searle and Penrose argue against strong AI, they are both convinced that intelligent and conscious systems do not need any non-physical component. This is different from the older idea which held that there was a non-physical component (spirit or soul) to a human, over and above the material body.

> Well, in one sense, of course, we are all machines. We can construe the stuff inside our heads as meat machines. . . . So, trivially, there are machines that can think.
>
> Searle: Minds, Machines and Science

Strong and Weak AI compared

How can we adjudicate between the strong and the weak AI positions?

Is there some experiment that could settle the dispute one way or another, by building a system and showing some of its behaviour?

At first sight, it seems that someone could show that strong AI is correct simply by building a system that had the properties of consciousness and so on that we want to build.

But how can we know the system is conscious, if all we can do is observe how it reacts? Perhaps it is just simulating consciousness?

So it's hard to see how this question can be answered by showing a system with particular properties, or looking at many systems. In fact, the question makes little difference to AI practice.

However, it's an important philosophical question, which is debated in philosophical terms. Searle claims to have an argument that shows that strong AI is wrong. This argument has been the centre of much controversy since Searle proposed it.

The Chinese Room

Searle compares a program that answers queries in Chinese to the following situation.

"Imagine you are locked in a room, and in this room are several baskets full of Chinese symbols. . . . you are given a rule book in English for manipulating these Chinese symbols. . . . the rule might say 'Take a squiggle-squiggle out of basket number one, and put it next to a squoggle-squoggle sign form basket number two.' Now suppose that some other Chinese symbols are passed into the room, and you are given further rules for passing Chinese symbols out of the room. . . . You are so good at manipulating the symbols, that very soon your answers are indistinguishable from those of a native Chinese speaker."

Searle in the Chinese room

Chinese room (ctd)

"There is no way you could learn any Chinese simply by manipulating these formal symbols.

Now the point of the story is simply this: by virtue of implementing a formal computer program from the point of view of an outside observer, you behave exactly as if you understood Chinese, but all the same you don't understand a word of Chinese. But if going though the appropriate computer program for understanding Chinese is not enough to give *you* an understanding of Chinese, then it is not enough to give *any other digital computer* an understanding of Chinese."

Syntax and Semantics

Searle backs up this idea by use of the ideas of *syntax* and *semantics* of languages (Chinese, in this case).

The *syntax* of a language tells us how sentences in the language are put together, what the vocabulary is, what the grammar is, etc. Procedures defined on the syntax of a language work just by looking at the shapes of the symbols involved, without consideration of their meaning.

The *semantics* of a language tells us what the meaning of the words and phrases is.

Searle's argument

Searle makes the following argument (here simplified), based on the Chinese room example.

Digital computers operate on symbols *purely in terms of the syntax*; but conscious understanding of a language also involves semantics, and (he claims) there is no way to get semantics just from syntax.

Therefore no digital computer can achieve conscious understanding, *solely* by running some program. (As we saw earlier, some systems achieve understanding, but they are not *just* digital computers.)

Objections to Searle

There have been many objections made to this argument. For example, Daniel Dennett in "Consciousness Explained" supports the "systems reply". This suggests that, even though the person in the room does not have understanding, the whole system (with the instructions, baskets etc) *does* understand.

Dennett appeals to "levels of understanding" that are needed to organise complex AI systems. The rules needed for the Chinese room would have to be very complex, and themselves organised in various levels. Though we don't want to attribute understanding to low levels (eg individual neurons in the brain), we *can* (says Dennett) attribute understanding to the whole system.

Of course, Searle has a counter-argument to this (and so it continues . . .).

Behaviourism again

Notice that Searle does not think that the internal state of his processing system is important when assessing whether there is any understanding (his description of computation makes no mention of "state").

Dennett thinks such a system would have to have a rich internal architecture (eg, not just be a very large look-up table), and puts the stress on the multi-layer system organisation.

From the behaviourist point of view, the distinction that Dennett is making would not be important; if an algorithm for understanding Chinese exists, described in a state-based way, then we would expect that some ways of showing the understanding behaviour involve understanding, and others do not.

Summary

- *Behaviourism*: just pay attention to externally observable behaviour
- Two positions on the possibility of building AI machines with understanding or consciousness:
  – **Strong AI**: any execution of an appropriate algorithm is enough;
  – **Weak AI**: all we can do is simulate these mental states.
- We also saw Searle's *Chinese room* argument, and a counter-argument.