Today

Some philosophical issues about AI

- Mind and Body
- Dualism, materialism
- Free will

Mind and Body

One of the oldest philosophical problems is that of the relationship between human minds and the physical universe.

The topic has been enlivened recently with the possibility of building systems that support artificial "mental" processing. Today we consider a couple of the traditional answers, and how they relate to AI systems.

See Searle's "Minds, Brains and Science" for one presentation of this area.

Mental and Physical Domains

The problem is this:
when describing people as intelligent agents, we attribute to them things like *consciousness, goals, beliefs, rationality* . . . .

Yet when we look at a description of a physical system (*eg* a human body), we find a description of a distribution of matter through space and time, that evolves according to physical law.

These two *levels of description* are very different.

What do they have to do with each other?

Physical States

It's useful to be able to characterise what's going on in a system in terms of the *state* of the system.

For a physical system (*eg* a mechanical clock), the state can be described in terms of a small, number of values, say the position of the hands of the clock, and the tautness of the spring. Once we know these values (assuming the clock is not broken in any way), we can tell how the clock will behave.

For a human brain, the description would be much more complicated. Other physical states are, for example:

$$\begin{array}{l} \textit{being upright} \\ \textit{having a temperature of } 35 \textit{ degrees } C \\ \textit{accelerating at } 9.81\ ms^{-2} \end{array}$$

Mental States

There is an everyday language that we use to describe the mental states of people (being *happy, cold, disappointed* . . .).

We don't have complete descriptions here, nor can we use these to predict very accurately future mental functioning; but this forms an important part of our understanding of how people are motivated and behave.

This everyday understanding of human behaviour in terms of mental states is known as *folk psychology*.

Mental or Physical?

- on the top of Arthur's seat ?
- in pain ?
- emitting radio waves at 250 kHz ?
- having light at the "red" wavelength impacting the retina ?
- seeing the colour red ?

Mind/Body again

We can restate the mind-body problem in terms of state as follows:

What is the relation between mental and physical states?

Dualism

This position says that there are two kinds of object in the world:

- *physical* objects (tables, ants, clouds . . .)
- *mental* objects (minds, thoughts . . .)

Dualism ctd

Since these are different sorts of object, they are expected to behave in different ways. For example, physical objects have some position in space; mental objects would not need to have a physical location.

The most famous advocate of this position was the French philosopher René Descartes (1596–1650).

Dualism (ctd)

Since mental and physical states are on this view states *of different objects*, then it's not surprising that they are described in such different ways.

For Descartes, the body was a machine which in itself had no feeling or purpose, but which could respond to events by reflex action. The body and the mind thus have two separate existences (this is where the name "dualism" comes from).

Understanding not in the brain

Descartes thought that understanding belonged to the mind ("soul") and not the brain:

> After having thus considered all the functions which pertain to the body alone, it is easy to recognise that there is nothing in us which ought to contribute to the soul, excepting our thoughts, which are mainly of two sorts, the one being the action of the soul, and the other its passions.
>
> (*Les Passions de l'Ame*)

Linking Mind and Body

If the mental and physical realms are separate, how do they influence each other?

Sensory input from the physical world is detected by the body, via physical means that are better understood now than in Descartes's time.

This affects the mind, which in turn can result in action of the body.

Descartes had an idea of how to explain this interaction, but it is seen as one of the weak points of his version of dualism.

An argument for Dualism

We say a system is *deterministic* if its future evolution is unique, given its present state.

One argument in favour of dualism is as follows:

> Since the mental and physical realms are distinct, they evolve in different ways. Thus the physical system might evolve deterministically, while some choice is available in the mental realm. This choice is necessary for humans to have free will. So we should adopt dualism.

As stated here, this is a bad argument, with several unstated steps, each of which needs justification.

Arguments against Dualism

Dualism has fallen out of favour as more scientific understanding of the functioning of the human body has appeared.

Present day physics no longer suggests that the universe evolves deterministically, so there is less force in the argument that a separate mental realm is needed for free-will.

The lack of any plausible account of how the mental and the physical interact is another reason to reject dualism.

Materialism

The opposite of dualism is called *monism* – it says that there is only one kind of object in the world, and not two.

Usually it is the physical, material domain that is used. The claim that mental states are in fact states of material objects is called *materialism*. It says that

> mental states are supported by their material realisation

e.g. via the states of the physical neurones of the brain.

Suppose that we know what state someone's brain is in when they are disappointed. Then "being disappointed" just means having a brain in that state.

Materialism ctd

An early version of this position is given by the 19th century naturalist Romanes. He wrote in 1885:

> We have only to suppose that the antithesis between mind and motion — subject and object — is itself phenomenal or apparent: not absolute or real. We have only to suppose that the seeming duality is relative to our modes of apprehension: and, therefore, that any change taking place in the mind, and any corresponding change taking place in the brain, are really not two changes but one change.
>
> *Mind and Motion*

Romanes is concerned not only to have a simpler account (just having one sort of object is simpler than having two interacting domains), he also wants to explain how we can have such different views of mental events, depending on whether we look at a brain scan or whether we ask how our own thoughts seem to us.

He suggests we have two different ways of perceiving and representing the same thing.

Note that this raises the possibility of experimental evidence: if the mental events just are the physical ones, then we can compare the time of experienced mental events with the time of their observed physical correlate.

Materialism: for

Materialism is appealing if we think that there is nothing more in the world than the material objects described by physics.

Our growing knowledge from neurophysiology about the details of the working of the brain make more aspects of mental life explicable in terms of brain functioning. We also know that damage to the brain results in changes to mental functioning, in a systematic way.

So this is an obvious scientific position to take.

Materialism: against

What are the objections to Materialism?

For one thing, since brain states are localised, it should follow that their mental states are localised too (so the pain I feel is situated just above my left ear, perhaps). This is bizarre . . .

The larger problem is that of accounting for *qualia*:
– what is it like to "see red"?
– what if colour experiences are swapped?

Free Will

Although materialism is largely accepted by present-day science, it looks as though it gives rise to problems for the notion of the free-will of people.

Recall: a system is *deterministic* if its future evolution is unique, given its present state. This means that for any initial state, there is only one possible sequence of states the system can follow.

An agent with *free will* should be able, when several courses of action are available, to select any one of them — compare the notion of an *autonomous* artificial agent.

Free will matters since it goes along with having moral rights and duties – for example, someone who does some act under hypnosis will not be thought of as responsible for that act.

Making decisions

> It has seemed very important to demonstrate that we are not just acting out our destinies but somehow choosing our own courses, *making* decisions—not just having "decisions" occur in us.
>
> Dennett, "Elbow Room"

Can artificial agents really be autonomous and make their own decisions?
What (if anything) is different between an artificial agent and a human here?

An argument

So, a problem arises if we accept materialism; we can then argue:

1. The mental state of a person depends on the physical state.

2. The physical system evolves in a deterministic way, for a given environment and initial state.

3. Therefore the mental state of a person is also determined by the environment (and the initial mental state).

4. Therefore we have no free will.

What's wrong?

We need to check the steps, and also whether the terms are used *consistently*.

[1] – let's accept this.

[2] – this is a claim about the physical world.

- This was held by physics up to 1910.
- Present physics is *not* deterministic in this way.
- But this gives randomness, not choice . . .
- We would prefer not to depend on basic physics for our free will!

What's wrong (ctd)

It looks as though [3] follows by simple logic. But let us look at this more closely.

Even if mental states depend on physical states, we still have two different *description languages*, one for mental notions, and the other for physical. Can we really go back and forth between the two levels of description in the way the argument suggests? Let's look at an example of a computer system where not all properties of one level of description are inherited by another.

Random Number Generators

A random number generator is a computer program which outputs a number, usually in a given range, at random each time it is called.

The essential property is that its behaviour should be *unpredictable* by the user. However, it is usually written in a programming language and run on a electronic computer – it is *predictable* from step to step at the execution level, and at the code level.

The art is to design a procedure whose inner workings are incredibly hard to reproduce without another computer, and whose output is not statistically biased. Typical techniques are: use the least significant digits of the computer's inner clock; divide a big number by a small one and use the remainder; or use the decimal expansion of π.

Predictability

Thus the same process can be described as predictable at one level, but unpredictable at another: the random number generator *is* predictable on the implementation level, but from the point of view of the programmer who uses it, it is designed to be *unpredictable*.

So *predictability* may not be inherited from level to level (this is different from determinism, though). The deterministic argument can be challenged at the transition from statement [2] to [3]. It now seems reasonable to describe the process of decision making as deterministic at the neural level, but involving free choice at the system level.

Mind, Body, Machine?

Let's try to relate this discussion to AI systems.

What corresponds to the mind and the body, in the case of an AI system? The physical machine corresponds to the body here. What corresponds to the mind?

We can describe the running system by the evolution of its physical state. Can such a system embody mental processes?

We follow this question in the next lecture.

Summary

- We looked at the problem of relating *mental states* to the *physical states* of an embodied agent.
- Two traditional answers involve saying that there are two separate realms in the world (dualism), and that mental and physical states are in fact the same thing (identity theory).
- Free-will and autonomous artificial systems.