Introduction to Philosophy

Antoni Diller

1 December 2011

1 Introduction

There are many different kinds of philosophy. For example, there is Indian philosophy (which has its roots in the ancient body of oral tradition known as the Veda), Islamic philosophy, Japanese philosophy and Chinese philosophy. However, none of these have had any real influence on the scientific tradition in which cognitive science has grown up. That scientific tradition is Western science and Western science is inextricably linked with Western philosophy. Wherever in the world today people study cognitive science or AI, they do so in the context of Western science and philosophy.

Western philosophy started with the Greeks about two and a half thousand years ago. Traditionally, the first philosopher is said to have been Thales of Miletus who lived in the sixth century BC. There have been many different philosophical schools and movements over the centuries and, currently, there are several different philosophical traditions. I shall focus on two that are particularly relevant to cognitive science and they are *analytical philosophy* and *critical rationalism*.

Analytical philosophy has been around for about a hundred years. It has its origins in the work of Gottlob Frege, Bertrand Russell and the logical positivists of the Vienna Circle. Its greatest exponent was Ludwig Wittgenstein. Critical rationalism is the name given to the philosophy of Karl Popper. He had many students who developed and criticised his ideas and, although they would not call themselves critical rationalists, they have more in common with Popper than with analytical philosophy. These students included Imre Lakatos, Paul Feyerabend, William Warren Bartley, III, David Miller, Ian Jarvie and Joseph Agassi. I will say more about these two philosophical movements later, but first I want to try and give you a feel for philosophy.

Many people have a very low opinion of philosophy. They think that philosophers occupy themselves with unanswerable pseudo-problems. The sociologist David Bloor writes (*Knowledge and Social Imagery* (1976), p. 45):

> To ask questions of the sort which philosophers address to themselves is usually to paralyse the mind.

Furthermore, many scientists think that the writings of philosophers are void of content. The philosopher Ludwig Wittgenstein does not help the cause of philosophy when he writes (*Philosophical Investigations* (1953), p. 261):

> So in the end when one is doing philosophy one gets to the point where one would like just to emit an inarticulate sound.

I hope to convince you that philosophy is important and inescapable. To begin with I will illustrate how philosophical problems differ from mathematical and scientific ones by giving some examples of different sorts of problem:

- mathematical problems:
  - Is there a largest prime number?
  - Do non-zero natural numbers x, y, z exist such that $x^n + y^n = z^n$, for any $n > 2$?
  - What is the relationship that holds between the faces, vertices and edges of a regular polyhedron?
- scientific problems:
  - What causes BSE (mad cow disease)?
  - How do the planets move?
  - Is there life on Mars?
- philosophical problems:
  - Are other people conscious?
  - Am I a brain in a vat?
  - Am I living in a computer-generated world?
  - Could an android be alive?
  - Could computers have emotions?

There are well-understood ways of tackling mathematical problems, involving various methods of proof and the construction of counter-examples, and there are well-established techniques for solving scientific problems, involving experimentation and other methods, but philosophical problems are not amenable to such ways of solving problems. That is one reason why so much of this module is taken up with methodology. I am trying to equip you to solve philosophical problems for yourself. As the great philosopher Immanuel Kant said, ‘You will not learn form me philosophy, but how to philosophize, not thoughts to repeat, but how to think.’ (This is quoted from Brenda Almond, *The Philosophical Quest* (1988), p. 11.)

2 Agreement or Diversity

The philosopher René Descartes wrote (*Discourse on Method* (1637), section 2):

> There is nothing so absurd or incredible that it has not been asserted by one philosopher or another.

He thought that this was a bad thing and sought to devise a method whereby certain truth could be achieved by everyone. The goal of certainty has long since been given up by philosophers, but some analytical philosophers still think that diversity is a bad thing. For example, the philosopher Michael Dummett wrote (*The Logical Basis of Metaphysics* (1991), p. 19):

> Philosophy would interest me much less if I did not think it possible for us eventually to attain generally agreed answers to the great metaphysical questions . . . .

I hope to convince you, however, that the proliferation of views is a good thing.

3 Why Study Philosophy

- Everyone has got a philosophy.
- Your philosophy influences many of the important things you do.
- Most people's philosophies are not worth very much.
- Your philosophy can be improved by criticism.

4 The Bucket Theory of the Mind

To illustrate the importance of philosophy I will give an example of a bad philosophical theory, empiricism, and how it has distorted some work in science. Whitehead explains *empiricism* as follows (*Adventures of Ideas* (1933), p. 228):

> [All] knowledge is derived from, and verified by, direct intuitive observation.

Empiricists view the mind as being a bucket. (See Fig. 1, which is based on Popper, *Objective Knowledge* (1975), p. 61, Fig. 3.) Popper elaborates the bucket theory of the mind as follows [6, p. 60]:

> [The] commonsense theory of commonsense knowledge is a naïve muddle. Yet it has provided the foundation on which even the most recent philosophical theories of knowledge are erected.
>
> The commonsense theory is simple. If you or I wish to know something not yet known about the world, we have to open our eyes and look around. And we have to open our ears and listen to noises, and especially to those made by other people. Thus our various senses are our *sources of knowledge*—the sources of the entries into our minds.
>
> I have often called this the bucket theory of the mind.

The empiricist view of science says that research begins with the collection of facts or data and only after a sufficient number of these have been gathered should the scientist begin devising a theory that explains them. This view of how knowledge is created

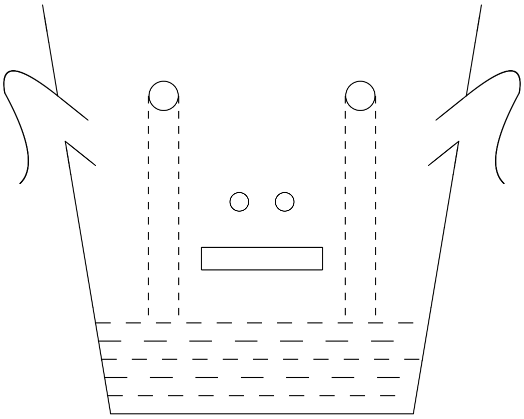

Figure 1: The bucket theory of the mind.

has distorted work in many sciences. I will give several examples in the course of this module. Here, I will give an example of how research in archaeology was hampered for many years because some prominent archaeologists accepted empiricism.

One of the things that archaeologists study is rock art produced by pre-historic humans. For many decades archaeologists have been trying to work out why pre-historic humans produced drawings on rock. One prominent archaeologist, Harald Pager, was influenced by empiricism and spent many years collecting facts about rock art. He also got other archaeologists to collect such data. One of those he influenced was David Lewis-Williams who spent years collecting facts about rock art. In 1975, for example, he and his colleagues recorded 2,361 rock-art images in 38 rock shelters in South Africa, but he had been collecting such information for at least the previous three years. Finally, it dawned on him that he was getting no nearer to understanding the significance of rock art. Many years later he wrote, in his book *A Cosmos in Stone* (2002), p. 49:

> I would not advocate quantitative techniques in rock art research unless important questions can be formulated in such a way that they can be answered numerically . . . . [Harald] Pager's 1971 book *Ndedema* contains an 80-page inventory that gives numerical descriptions of 3,909 individual rock paintings. As far as I am aware, no one has used this compilation to answer any questions that Pager himself did not pose. This is a cautionary tale for those who gather 'objective' numerical data in the hope that others will be able to use them.

Lewis-Williams went on to devise the so-called neuropsychological theory of rock art which states that rock-art paintings are an attempt to copy the images that a shaman sees while he is under the influence of an hallucinogenic substance. I will say more about empiricism and how research in science should proceed in a later lecture.

The above has been a Popperian rationale of the value of philosophy. Analytical philosophers see the function of philosophy very differently.

5 Analytical Philosophy and Cognitive Science

5.1 Introduction

Wittgenstein was of one the most influential analytical philosophers and he did not think that philosophical problems were either real or genuine. He thought that philosophical problems arose through the misuse of language. For him, the function of philosophy was to remove the conceptual confusions that made people think philosophical problems were real. He famously wrote (*Philosophical Investigations* (1953), p. 232):

> The confusion and barrenness of psychology is not to be explained by calling it a "young science"; its state is not comparable with that of physics, for instance, in its beginnings. (Rather with that of certain branches of mathematics. Set theory.) For in psychology there are experimental methods and *conceptual confusion.* (As in the other case conceptual confusion and methods of proof.)
>
> The existence of the experimental method makes us think we have the means of solving the problems which trouble us; though problem and method pass one another by.

Analytical philosophers see the theory of meaning as forming the foundation of philosophy. They, therefore, spend much of their time thinking about meaning and studying language. By doing this they hope to remove the muddles that people get into when they think about "philosophical" issues. Some, like Dummett, see the construction of a generally agreed theory of meaning as being the means of resolving traditional philosophical disputes.

5.2 Searle's Chinese-room Argument

John Searle is a prominent analytical philosopher and he devised the Chinese-room argument which he claims refutes strong AI, that is to say, the view that holds that 'the mind is a computer program and the mind is to the brain as the program is to the hardware' [7, p. 546]. I quote his argument because it is much discussed by cognitive scientists and also because it well illustrates the analytical approach to "philosophical" issues.

> Imagine that I, a non-Chinese speaker, am locked in a room with a lot of Chinese symbols in boxes. I am given an instruction book in English for matching Chinese symbols with other Chinese symbols and for giving back bunches of Chinese symbols in response to bunches of Chinese symbols put into the room through a small window. Unknown to me, the symbols put in through the window are called questions. The symbols I give back are

> called answers to the questions. The boxes of symbols I have are called a database, and the instruction book in English is called a program. The people who give me questions and designed the instruction book are called programmers, and I am called the computer. We imagine that I get so good at shuffling the symbols, and the programmers get so good at writing the program, that eventually my 'answers' to the 'questions' are indistinguishable from those of a native Chinese speaker. I pass the TURING test for understanding Chinese. But all the same, I don't understand a word of Chinese and—this is the point of the parable—if I don't understand Chinese on the basis of implementing the program for understanding Chinese, then neither does any digital computer solely on that basis because no digital computer has anything that I do not have.

The Turing test that Searle mentions is Turing's suggestion [8] that, if we cannot tell the difference between talking to a machine and talking to a human, after communicating with a machine, not knowing it to be a machine, for an extended period of time, then we should call the machine intelligent. Searle's argument has many faults. The most serious is that Searle says that the person in the room passes the Turing test ('I pass the TURING test for understanding Chinese'—lines 5–6 from the bottom of the quotation), but the person in the room does *not* pass the Turing test. What passes the Turing test is the conglomerate formed out of the program, the database and the computer, not just the computer by itself. (Note that the person in the room corresponds to a computer.)

6 How to Detect an Android

6.1 Some Questions and an Argument

Many philosophers, cognitive scientists, roboticists and people interested in AI discuss the following sorts of question:

- Is it possible to design and build an undetectable android?
- Will androids ever come to have human emotions?
- Will androids ever be conscious in exactly the same way that human beings are conscious?

My approach to these questions is influenced by Popper's philosophy and is very different from how analytical philosophers deal with such questions. I will argue that it is impossible to build an undetectable android, but this argument will not involve dubious claims about meaning or conceptual confusion. The following is a summary of this argument that I have published elsewhere [1].

In order to get the android to behave like a human being the people designing the android will have to equip the android with various programs that produce human-like behaviour. They will make use of various psychological and other theories to write these programs. Thus, a large number of theories will be incorporated or embedded in the android in order to try and get him to produce the full repertoire of human behaviour. To be potentially undetectable the android would have to be able, for example, to express emotions appropriate to the various circumstances in which he finds himself. He would also have to be able to communicate verbally with human beings about a wide range of topics and act appropriately in different settings such as restaurants, banks, pubs, post offices, libraries, churches, cinemas, casinos and so on. He would also have to be able to do a large number of other things. To make the argument easier to follow I will concentrate on just one type of behaviour. I will restrict my discussion to behaviour which is associated with emotional states in human beings. Furthermore, I will assume that a single theory is responsible for producing emotion-related behaviour in the android. In reality, several complementary (and, hopefully, mutually consistent) theories would probably be used, but for the purposes of my argument these can be thought of as having been combined into a single theory. It should be noted that I am not here concerned with issues relating to whether or not the android could experience anxiety, say, in exactly the same way that a person does. Nothing that I say depends on the felt quality of different emotions. I am solely concerned with the external manifestations of some emotions. In the case of extreme anxiety, for example, these might include trembling, shortness of breath, excessive sweating and difficulty in swallowing. In addition, I am neither assuming that every emotional state experienced by an adult has some distinctive behavioural counterpart nor am I assuming that all expressions of emotion are sincere. To be undetectable the android would have to be able to hide his emotions on some occasions and also to falsely express emotions that he was not experiencing on other occasions. It should be noted that with suitable changes the argument that I am presenting would apply to all the different kinds of behaviour in the android's repertoire and not only to emotion-related behaviour.

Unfortunately, all scientific theories are false. Therefore, the theories embedded in the android which produce emotion-related behaviour are false and this enables the android to be detected. The claim that all scientific theories are false is the most controversial premise that my argument makes use of and I will, therefore, return to it later and discuss it more fully. Before that, however, I will show how an android can be detected if you accept, for the sake of argument, that all scientific theories are false.

The behaviours which allow us to detect the android are those which falsify the theories embedded in him. (In general, different behaviours would falsify different theories.) Such falsifying behaviour cannot be exhibited by the android because he can only behave in accordance with the theories incorporated in him that produce his behaviour. I will present an analogy in order to clarify this point:

> An orrery is a clockwork model of the solar system. An orrery could be built to illustrate Ptolemy's celestial mechanics with the Earth at the centre and the sun and other planets moving around it in orbits that are produced by combining several circular motions or it could be built to illustrate Copernicus's theory in which the orbits of the planets are again obtained

> by combining several circular motions but now the Sun is at the centre of the system. Yet another orrery could be built to illustrate Kepler's celestial mechanics. In this model the planets would move in ellipses around the Sun which would lie at one of the foci of each ellipse. In practice, an orrery cannot be built which is an exact scale model of the solar system. (One reason why this is the case is because the highest common factor of the mean distances of the planets from one another is very small in comparison with the mean distance of the furthest planet from the Sun.) A computer model could, however, be constructed. Imagine such a model which was built to illustrate Ptolemy's celestial mechanics. The behaviour of the model could not deviate from that described by Ptolemy's theory. No matter how many observations we took of the miniature solar system they would always be in conformity with that theory. It would be impossible for any observations of that model to falsify Ptolemy's theory. We know, however, that the behaviour of the actual solar system is different from what we would expect on the basis of Ptolemy's theory. The behaviour of the solar system that falsified Ptolemy's astronomy could not be produced by the computer model.

In this analogy the computer model (or orrery) corresponds to an android and the behaviour of the model corresponds to the android's emotion-related behaviour, the real solar system corresponds to a human being and its behaviour corresponds to human emotion-related behaviour. Furthermore, Ptolemy's celestial mechanics corresponds to the theory of emotion that was used in order to produce the android's emotion-related behaviour. There are discrepancies between the behaviour of the android and human behaviour in the same circumstances just as there are discrepancies between the behaviour of the model and the behaviour of the solar system. In the case of emotion-related behaviour these discrepancies allow us to tell androids apart from humans. Let us say that the design team uses a theory of emotion T to get the android to produce human emotion-related behaviour. Theory T is a model of human emotion which we can assume is produced by a human being's social brain P. (For convenience, I use the phrase 'social brain' to refer to those aspects of a human being or android that produce its emotion-related behaviour.) If T is used in the design and manufacture of an android's social brain C, then that android will display emotion-related behaviour which has been produced by C which instantiates theory T in its design. Thus, whereas human emotions are produced by P, android emotions are produced by C which was designed using a theory T which is a model of P. In these circumstances it would be possible for human behaviour and emotional response to falsify theory T, but it would not be possible for android behaviour and emotional response to falsify T. As a theory of human emotion T is a falsifiable empirical theory, whereas as a theory of android emotion produced by component C it is an unfalsifiable, non-empirical theory. It would be possible for theory T to be falsified by experiments involving human beings, but it would be impossible for android behaviour to falsify the theory. Because all scientific theories are false, we know that T is false and some human behaviour does actually falsify it. This means that the difference between human emotional response and android emotional response can be established by any human behaviour that falsifies theory T, because an android could not exhibit such behaviour. Thus, there is a way of detecting the presence of androids in human society.

Based on the considerations presented the following method is proposed as a means of deciding whether or not a test subject is an android:

(1) Find out what theory of emotion T has been instantiated in the androids that live amongst us.

(2) Find out what human behaviour B falsifies T.

(3) Put the test subject in a situation where a human would exhibit behaviour B.

(4) If the subject exhibits B, it is human. If the subject fails to exhibit B, it is an android.

6.2 All Scientific Theories are False

The most controversial premise in the above argument is the claim that all scientific theories are false. To try and convince you of the truth of this claim I begin by mentioning that the history of science is littered with false scientific theories. Here are a few examples:

- the theory of spontaneous generation
- the caloric theory of heat
- Aristotelian mechanics
- the theory that light moves infinitely fast
- Ptolemaic astronomy
- the theory that the chemical atom is indivisible
- the phlogiston theory

Several philosophers have reflected on the fact that the history of science is full of false theories and some have come up with an argument that they call the *pessimistic induction.* Newton-Smith [4, p. 183] formulates it as follows:

> Past theories have turned out to be false, and since there is no good reason to make an exception in favour of our currently most cherished theories, we ought to conclude that all theories which have been or will be propounded are strictly speaking false.

Newton-Smith even goes so far as to say [4, p. 14], ‘Indeed the evidence might even be held to support the conclusion that no [scientific] theory that will ever be discovered by the human race is strictly speaking true.’ In order to avoid paradox it is important to emphasise that the conclusion of the pessimistic induction, namely the claim that all scientific theories are false, is not itself a scientific theory of the same sort or level as the theories it applies to.

Inductive arguments are not deductively valid and so it is possible for their conclusions to be false even when all of their premises are true. They are, however, both in science and in philosophy, often extremely useful as heuristics. In science, for example, they are frequently very helpful in suggesting universal scientific theories that the scientific community then tries to confirm or refute. It is those attempts and their interpretation that confer plausibility or implausibility on the theories in question. Similarly, in philosophy, inductive arguments, such as the pessimistic induction, are sometimes exceedingly fruitful in indicating general empirical theories. These are then subjected to a searching evaluation and it is that assessment and its interpretation that helps us decide on the truth or falsity of the universal theories in question. Thus, the conclusion of the pessimistic induction, namely that all general scientific theories are false, has not been conclusively established, but the pessimistic induction does provide strong *prima facie* evidence in its favour. Much more could be said about the view that all scientific theories are false, but my main concern is not to present an exhaustive account of the debate that the pessimistic induction has given rise to. If the reader is undecided about the correctness of the view that all scientific theories are false, then he may still be interested in the consequence that I show follows logically from this position and especially in the following conditional sentence: If all scientific theories are false, then an undetectable android cannot be made. Someone who is in two minds about the claim that all scientific theories are false is still likely to find the contrapositive of the previous conditional sentence downright counter-intuitive: If an undetectable android can be made, then some scientific theories are true.

Although many people think that is is reasonable to believe that all scientific theories are false, this position does not have disastrous consequences for science. If we look at a chronological series of theories in any given area of science, what we usually find is that later theories in the series are better than earlier ones. For example, in celestial mechanics, Einstein's theory is more accurate than Newton's and Newton's is an improvement on Kepler's. Kepler's theory, in turn, is much better than Copernicus's, which itself was a great improvement on Ptolemy's. Although all these theories are false, the later ones are better than the earlier ones. One way in which the later theories are better is that they give rise to more precise predictions of astronomical phenomena. Furthermore, just because a theory is false it does not follow that it is useless. Newton's theory, for example, is false, but it was used by NASA scientists in order to plot the trajectories of rockets sent to the moon and those rockets reached their destination. Newton's theory is false, but it is a good approximation to the truth.

7 Some Traditional Philosophical Problems

7.1 Introduction

In this section I want to look briefly at some philosophical problems that are directly relevant to cognitive science.

7.2 The Problem of Personal Identity

The traditional problem of personal identity can be formulated as follows:

> People change as they grow older. Let us consider Ludwig Wittgenstein as an example. He was born on the 26th of April 1889 in Vienna. He grew up into a child and then he became an adult and he died an old man on the 29th of April 1951 in Cambridge. Even though he underwent many physical and mental transformations he remained the same human being. What makes the old man the same human being as the baby? Are they the same person?

Philosophers have put forward several theories of personal identity:

- immaterial soul (Plato, Descartes)
- memory continuity (Locke)
- psychological continuity
- spatio-temporal continuity
- physical continuity (of brain and/or body)

Few, if any, cognitive scientists would accept that souls exist. However, if you believe in the existence of souls, it is easy to prove that computers can be conscious—especially is you also believe in re-incarnation. The Dalai Lama puts the matter as follows:

> I can't totally rule out the possibility that, if all the external conditions and the karmic action were there, a stream of consciousness might actually enter a computer. There is a possibility that a scientist who is very much involved his whole life with computers, then the next life he would be reborn in a computer.

(This is adapted from Hayward and Varela (eds.), *Gentle Bridges: Conversations with the Dalai Lama on the Sciences of the Mind* (1992), p. 152.)

7.3 The Problem of Other Minds

The philosophical problem of other minds can be expressed as follows:

> How can I know that other people have conscious experiences that are similar to the ones that I have? How can I be sure that other people have any conscious experiences at all?

7.4 The Body-mind Problem

According to Popper (*Realism and the Aim of Science* (1983), p. 103), the body-mind problem is 'the problem of the immensely intricate physiological influences (of drugs, say) upon our mental state, and *vice versa*, of mental influences (of the realization of dangers, say) upon our physiological state.' I will look at the body-mind problem in more detail in the next lecture.

References

[1] Antoni Diller. Detecting androids. *Philosophy Now*, (25):26–28, Winter 1999/2000.

[2] Michael Dummett. *The Logical Basis of Metaphysics.* Duckworth, London, 1991.

[3] David Lewis-Williams. *A Cosmos in Stone: Interpreting Religion and Society Through Rock Art.* AltaMira Press, Walnut Creek (CA), 2002.

[4] W. H. Newton-Smith. *The Rationality of Science.* International Library of Philosophy, editor: Ted Honderich. Routledge & Kegan Paul, Boston, London and Henley, 1981.

[5] Harald Pager. *Ndedema: A Documentation of the Rock Paintings of the Ndedema Gorge.* Akademische Druck, Graz, 1971.

[6] Karl Raimund Popper. *Objective Knowledge: An Evolutionary Approach.* Oxford University Press, London, 1975. Originally published in 1972.

[7] John R. Searle. Searle, John R. In Samuel Guttenplan, editor, *A Companion to the Philosophy of Mind*, volume 5 of *Blackwell Companions to Philosophy*, pages 544–550. Basil Blackwell, Oxford, paperback edition, 1995.

[8] Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, October 1950.

[9] Alred North Whitehead. *Adventures of Ideas.* Cambridge University Press, Cambridge, 1933.

[10] Ludwig Wittgenstein. *Philosophical Investigations.* Basil Blackwell, Oxford, 1953.