PHILOSOPHICAL FOUNDATIONS OF ARTIFICIAL INTELLIGENCE

A course for second year AI students

Aaron Sloman

INTRODUCTION

There are many ways of teaching philosophy. A standard way is to ensure that students know what various philosophers have thought and written.

This course is too short for that. So I shall aim mainly to try to ensure that students learn two main kinds of things.

(a) A family of new concepts that are useful for formulating and discussing philosophical questions.

(b) How to *do* philosophy. That is very hard to teach. I don't really know how communicate it except by doing philosophy and helping students to do it, hoping that they will somehow pick up the techniques by imitation and practice. I tried to outline some of the techniques in my 1978 book *The Computer Revolution in Philosophy*, which is now out of print. However *describing* the process of philosophising does not seem to be good a way of communicating what it is about to those who have not yet learnt to do it.

To that extent, learning to do philosophy is something like learning to play the violin, or ride a bicycle. It's very easy to play out of tune, or lose your balance.

As this is a single module course you are expected to do on average about 6.5 hours a week of work on the course, including all the time spent in lectures or classes, reading, writing notes, thinking, etc. A significant amount of that time should be spent in the main library finding things out for yourself, e.g. initially by looking for encyclopaedias and dictionaries of philosophy, so that you learn where to find things out quickly, reading books that give overviews, and later on reading recent articles which analyse some of the problems we'll discuss.

You may find it useful to start with an introduction to some fairly 'traditional' ways of doing philosophy. E.g. try to find one of the introductory books listed below, e.g. by Campbell, Hospers, Magee (on Popper), Mitchell, or one of the collections of readings in philosophy by Edwards & Pap, Goldman or Lycan. There are *very many* more books that give introductory overviews, and you will find others by looking in the library for introductions to philosophy, to metaphysics, to epistemology, to philosophy of mind, or philosophy of science. For the first two weeks of the course, simply try to do as much general reading as you can, and bring to the classes any questions you have about the arguments, or concepts, or theories you find.

The course is assessed by an essay to be submitted in the Summer term. In order to be able to write a good essay you will need practice. So each student will be expected to introduce a

discussion of one or two topics this term, and to write a sample essay on which you can get critical feedback. Details will be arranged later.

ASPECTS OF PHILOSOPHY

There are many different ways of dividing up philosophy. You may recall the question we set you at the beginning of your course here:

"Can a goldfish long for its mother?"

This raises many different aspects of philosophy:

Metaphysics: What kinds of things exist? (E.g. fish, material things, mental states, relationships.)

Epistemology (or theory of knowledge): What can we know and how do we know it? In particular how could we tell whether a goldfish does or does not long for its mother? Can we ever know about the contents of "other" minds? For that matter can we know about anything other than our own mind and its contents?
Note: The question "How can we know X?" has at least two very different interpretations. On one interpretation it asks about the causal history of our knowing: did we see the evidence, or hear it, read about it, or infer it from some experiment, use our own tests or belief what someone else said, etc. On the other interpretation it is a question regarding justification: is the means by which you came to know X *sufficient justification* for claiming that it is true, or that you know it is true?

Philosophy of mind: What are mental states and processes? What's the relationship between mind and body? Are certain material states *sufficient* to produce mental states? Can mind and matter interact causally? Could longing cause swimming? Can sadness cause weeping?

Ethics or moral philosophy: How should we think about the rights of a goldfish? Do we have the right to kill them? To cause them pain? To take them from their mothers?

Philosophy of science: If someone thinks it is a scientific question whether the goldfish feels pain, then we can ask what the difference is between science and other types of knowledge, or knowledge-seeking? What are: scientific theories? Explanations? Evidence? Can theories ever be proved, or refuted, and if so how? What's the relationship between the development of new concepts and the development of new theories?

Conceptual analysis: It soon becomes clear that we are not sure what question we are asking? What does it *mean* to say that a goldfish longs for something? What does it *mean* to say that it can think about its mother? Or that it has a mother? (Could a tree or a rock have a mother? What about a battle?) There are many concepts we use outside of doing philosophy, which are extremely difficult to analyse. Examples of such concepts are: *mind, matter, meaning, truth, causation, experience, freedom, goodness, concept. knowledge, explanation, science, intelligence, emotion,* and many more. A great deal of modern philosophy attacks old problems by showing that the questions were confused because the concepts used were full of muddles,

like “Where is the universe and which way is it moving?”

I believe that learning to do conceptual analysis is one of the most important aspects of learning to do philosophy: everything else hangs on it. E.g. people who are unclear about the concepts they use can argue at cross purposes, or flounder in interminable debates (for examples see most of the discussion in comp.ai.philosophy).

But it is also relevant to being a good scientist. One of the most spectacular examples was how Einstein’s attempts to analyse the our concept of simultaneity led to the special theory of relativity.

Last year you had a brief introduction to conceptual analysis when we talked about similarities and differences between: *embarrassment, shame, guilt, regret* and related concepts. To help you get back into doing philosophy have a go at trying to write down what these states have in common and how they differ.

There is a fairly terse introduction to conceptual analysis in Chapter 4 of The Computer Revolution in Philosophy (1978) now available online here

`http://www.cs.bham.ac.uk/research/cogaff/crp/chap4.html`

Note:
There are many variants on the question about a goldfish. For instance you could ask about other mental states, or about about other animals. Here are some examples (in each case add the the question: “What does this question mean?” “What techniques of enquiry or analysis are relevant to deciding whether one answer is better than another?”

1. Could a mouse desperately hope that her children will do well in life?
2. Could a tadpole hope that it will survive to be a frog and make more tadpoles?
3. Is a fly aware of my fly-swatter coming down to flatten it? (If not, why does it always escape?)
4. Is the fly afraid of being hit by the fly-swatter? (If not, why does if fly away?)

THE RELEVANCE OF AI TO PHILOSOPHY

It is fairly evident that philosophy is relevant to AI, e.g. helping to set its goals and clarify many of the concepts it uses, such as *intelligence, perception, learning, memory, understanding*, etc.

Equally, AI and Computer Science are relevant to philosophy because they provide a host of new concepts and forms of explanation, as well as raising new questions relevant to old philosophical problems, about metaphysics, about what we can know, about the relationship between mind and matter.

For example: what is a virtual machine? What’s the relationship between virtual machines and physical machines? Can virtual machines enter into causal relationships? Can a “software event” like the creation of a new data-structure (e.g. a new list), *cause* physical events to occur, or is it only physical things that can enter into causal relationships? Do computing systems have “emergent” properties? How do computational machines differ from previous sorts of

machines? What are machines? Are connectionist machines significantly different from symbol manipulating machines? What are symbol manipulations? Isn't changing the weight on a neural link a sort of symbol manipulation?

A PROVISIONAL PLAN FOR THE COURSE

It is impossible to plan a philosophy course without knowing the philosophical capabilities of the students well. So here is a provisional list of topics, which we may pursue in rough chronological order, though if appropriate we can change the order. In fact there is no *right* order for learning philosophical concepts and theories: like learning a new town you have to go round and round getting to know things better by learning their mutual relationships. So here's a possible sequence for the course.

(1) Philosophical concepts and jargon. Make the acquaintance of some terminology, e.g. notions like *epistemology, metaphysics, ontology, monism, dualism, reductionism, physicalism (materialism), behaviourism, phenomenalism, idealism, epiphenomenalism, interactionism, intentionality, "derivative" intentionality, concept, proposition, rationality, concept empiricism, knowledge empiricism, supervenience, the analytic/synthetic and empirical/apriori distinctions, deductive nomological explanations, the design stance, the intentional stance (Dennett),*, and various theories about the relation between mind and body.

(2) Computer science, AI and Philosophy. Try to get a feel for some of the philosophical problems raised by computing and AI. (E.g. What is a machine? What is computation? What's the relationship between computational processes and physical processes? Can a software event *cause* a physical event? What is intelligence? Can there be a behavioural criterion for intelligence? What's the status of the Turing test? Can computational processes support semantics? Are neural processes in some fundamental way different from processes on a digital computer?)

(3) Machines and intentionality. What does it mean to say that a machine refers to something, or understands something? Searle's chinese room argument, and replies to it. Causal theories of meaning. Alternatives to causal theories of meaning. Harnad's "Symbol Grounding Problem."

(4) Representations. What are they? What is their role in intelligence? How many different kinds are there? Does logic have a special role? What, if anything, is special about pictorial or diagrammatic reasoning?

(5) What sorts of machines could have machines? Are there some aspects of mind that are particularly difficult to accommodate in machines in general, or in computers in particular? Qualia? Pains and pleasures? Emotional states? Consciousness? Experience? How could this be settled? Can machines have what Haugeland calls "original intentionality" as opposed to "derivative intentionality"? The relevance of new architectures to new analyses of old concepts. What's the difference between simulation and replication? When is a simulation of a Y a Y?

(6) Freedom of the will and related concepts. What Minsky calls "Dumbell theories" (everything is either an A or a not-A and there's nothing in between) and what is wrong with them. What could it mean for a machine to have its own goals? What does it mean for us to have our

own goals? Do we have a kind of freedom machines could never have?

(7) The importance of architecture. When you have an architecture, it defines a collection of possible states and processes. (Think of how current theories of the architecture of matter define different kinds of stuff - different elements, different chemical compounds, etc. Compare how people previously thought of water, iron, air, etc.) How many of our ordinary mental concepts presuppose specific architectures? What sorts of architectures can support mental states and processes?

Other possible topics to be decided later. Maybe we should talk about consciousness. There are many other concepts of ordinary language that we could try to analyse and relate to the possibility of instantiation in machines, e.g. em sensory experience, learning, desire, pain, pleasure, emotion, personality.

REFERENCES

Philosophy is a very old subject and has spawned a vast and diverse array of books and journals from many cultures. Here is a tiny subset of pointers into that literature, including some traditional introductions to philosophical ideas and also some more recent inspired particularly by developments in AI. I have not yet read all of this myself. Some of it is included on the basis of recommendations from others.

1.BOOKS

Boden, M. (1990) *The Creative Mind* Abacus edition, 1992.

Margaret Boden (Ed) *The Philosophy of Artificial Intelligence* (ed.) (Oxford University Press,) 1990.

Churchland, P.M. (1984). *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind.* Cambridge MA: MIT Press.

Campbell, Keith *Body and Mind* Macmillan, 1970.
A useful little introduction to traditional philosophical theories of mind.

Dennett D.C. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology* Cambridge, MA: MIT Press.

D.C. Dennett, (1984) *Elbow Room: the varieties of free will worth wanting,* Oxford: The Clarendon Press,
Compare my notes on how to dispose of the free will issue
`http://www.cs.bham.ac.uk/research/cogaff/0-INDEX81-95.html#8`

D.C. Dennett, (1996) *Kinds of minds: towards an understanding of consciousness,* Weidenfeld and Nicholson, London.

Paul Edwards & Arthur Pap (eds) (1957) (and various new editions) *A Modern Introduction to Philosophy* Collier Macmillan, New York

Fetzer, J.H. (ed) *Epistemology and Cognition* Kluwer Academic, 1990,

Haugeland, John, (ed) *Mind Design: Philosophy, Psychology, Artificial Intelligence,* Bradford Books, MIT Press, 1981.

Hofstadter D.W. and Dennett D.C. (Eds.) (1981). *The Mind's I: Fantasies and Reflections on Self and Soul* Brighton: Harvester Press.

John Hospers (1973) *An Introduction to Philosophical Analysis* Routledge and Kegan Paul
(Many students have found this a very useful (if somewhat wordy) general introduction to philosophy.

Lycan, William G. (ed) *Mind And Cognition: A Reader* Oxford: Basil Blackwell, 1990.

Magee, Bryan, *Popper* Fontana Modern Masters Series.
(A very good short introduction to Karl Popper's views of knowledge and science.)

Mitchell, David, *An introduction to logic* Hutchinson, 1962.
(This is mainly an introduction to philosophical analysis of logical concepts rather than an introduction to formal logic.)

Russell, Bertrand *The Problems of Philosophy*
(An ancient paperback book)

Ryle, Gilbert *The Concept of Mind* Hutchinson 1949. (A seminal, yet still underrated book).

Searle J. (1984). *Minds, Brains and Science: The 1984 Reith Lectures* London: BBC Publications. (An attack on AI)

Sloman, Aaron (1978) *The Computer Revolution in Philosophy: Philosophy science and models of mind*, Harvester press.
(Out of print but now available online with some afterthoughts added:
`http://www.cs.bham.ac.uk/research/cogaff/crp/`

2. JOURNALS

There are many journals that include discussion of philosophical issues relevant to Computing, AI and Cognitive Science. Over the last ten years or so, standard philosophical journals have increasingly included such articles, including the following:

Mind

The British Journal for the Philosophy of Science

There are also now journals that are explicitly concerned with issues to do with mind, brain and AI, and include philosophical and non-philosophical articles.

Artificial Intelligence

Artificial Intelligence Review

Computational Intelligence

Behavioral and Brain Sciences

Cognitive Science

Minds and Machines

The Monist

New Ideas in Psychology

and many more.

3. CONFERENCE PROCEEDINGS

Several AI conferences include philosophical papers, including, for example, proceedings in the following series of regular conferences:

International Joint Conference on Artificial Intelligence (IJCAI)

American Association for Artificial Intelligence (AAAI)

European Conference on AI (ECAI)

And conferences on Artificial Life.

4. THE INTERNET

There are increasing numbers of World Wide Web sites, some of which include philosophical references or articles.

David Chalmers has an extensive online bibliography and pointers to many useful web sites:

```
http://www.u.arizona.edu/ chalmers/resources.html
```

A site with general philosophical information is

```
http://www-personal.monash.edu.au/ dey/phil/
```

Many more internet sites giving information about philosophy can be found by giving google some combination of these words and phrases: philosophy, epistemology, philosophy of mind, philosophy of language, philosophy of law, introduction tutorial, etc.

There are various philosophical articles included in the Birmingham CogAff Web directory, accessible as:

```
http://www.cs.bham.ac.uk/research/cogaff/
```

Some of my slide presentations including philosophical presentations are here

```
http://www.cs.bham.ac.uk/research/cogaff/misc/talks/
```

5. USENET NEWS GROUPS – THE INTERNET

There are several usenet groups that include philosophical discussion from time to time. Often the “signal to noise” ratio is not very high.

The easiest way to read news groups is to use Google’s groups facility:

```
http://www.google.com/grphp?hl=en&tab=wg&ie=UTF-8&oe=UTF-8
```

You can try these news groups:

```
comp.ai.philosophy
sci.philosophy.meta
sci.philosophy.tech
alt.consciousness
sci.cognitive
```

```
sci.psychology.consciousness
alt.consciousness
comp.ai
```

There is an “electronic” journal of AI that is accessed via Usenet, the “Journal of AI Research”. See the two groups:

`comp.ai.jair.announce`
: Includes announcements of new papers available on JAIR.

`comp.ai.jair.papers`
: Includes the actual papers, circulated in compressed, uuencoded form.
(Ask for help if you don’t know what that means.)