How AI relates to philosophy and in some ways IMPROVES ON PHILOSOPHY

Talk to first year AI students, University of Birmingham, December 2001. Accessible as talk 13 here
http://www.cs.bham.ac.uk/~axs/misc/talks/
(Talk 10 is What is Artificial intelligence.)
WHAT IS PHILOSOPHY?

The most general of all forms of enquiry, with a number of more specific spin-offs.

PHILOSOPHY INVESTIGATES:

- The most general questions about what exists:
  - Metaphysics and ontology:
    An attempt categorise the most general forms of reality and possibly to explain why reality is like that. E.g. Can mind exist independently of matter?

- The most general questions about questions and possible answers:
  - Epistemology:
    an attempt to characterise the nature of knowledge and to identify the kinds of knowledge that are possible and the ways of acquiring knowledge.
  - Theory of meaning:
    An attempt to clarify the nature of meaning and how it differs from nonsense.

- The most general questions about what ought to exist, or ought not to exist:
  - Ethics (moral philosophy) and aesthetics
    an attempt to distinguish what is good and what is bad, including what is good or bad in art. Meta-ethics investigates the nature of ethics.

Contrast a naive popular view of philosophy: a study of the meaning of life?
More specific areas of philosophy

Besides the very general branches of philosophy there are many sub-branches which combine the above three philosophical studies in focusing on a particular form of human activity: Philosophy of X.

Examples include:

– Philosophy of mind
– Philosophy of mathematics
– Philosophy of language
– Philosophy of science
– Philosophy of history
– Philosophy of economics
– Philosophy of biology
– Philosophy of psychology
– Philosophy of literature
– Philosophy of computation
– Philosophy of music
– Philosophy of sport ??
– Philosophy of ....
Philosophy and science

The methods of science are both:
– theoretical: building explanatory theories; and
– empirical: observing, measuring, testing, doing surveys, etc. to find out what sorts of objects, events and processes exist in the world.
(See chapter 2 of http://www.cs.bham.ac.uk/research/cogaff/crp/)

The methods of mathematics are:
– purely theoretical: exploring, analysing, comparing, and proving properties of abstract structures, irrespective of whether anything in the observable world corresponds to them.

The methods of philosophy are:
– Partly like mathematics: philosophers do not conduct observations, experiments, surveys, etc. – philosophy is mostly non-empirical.
– Partly like science: Philosophers are concerned with what actually exists and can be known. Often conceptual: Attempting to get clear about how our own concepts actually work, what they presuppose, what criteria they use.

As a branch of philosophy makes its questions more precise and its methods more powerful, it may spawn a new science, e.g. physics, astronomy, biology, psychology, linguistics, ...
Philosophy of mind

This has several different aspects:

- **Metaphysical and ontological topics**
  Questions about the nature of mind and the relation between mind and body, e.g. whether and how mental events can cause physical events or *vice versa*.

- **Epistemological topics (theory of knowledge)**
  Questions about we can know about the minds of others, and how we can acquire such knowledge.
  More subtle questions about what we can and cannot know about our own minds: e.g. do you know which rules of grammar you use?

- **Conceptual analysis**
  Analysis of the concepts we use in talking about our mental states and processes, e.g. ‘perceive’, ‘desire’, ‘think’, ‘plan’, ‘decide’, ‘enjoy’, ‘conscious’, ‘experience’ ...

- **Methodology**
  Investigation, comparison and evaluation of the various ways of studying human (and animal) minds, including the methods of psychology, neuroscience, social science, linguistics, philosophy, and AI.
AI can be seen as an extension of philosophy of mind:

– We can survey different possible kinds of minds by asking how we could design and implement them.  
  (So far AI has produced only very simple examples)
– We can clarify the relationship between mind and body by treating it as an example of a relationship we understand better: the relationship between virtual machines (running programs) and physical machines (computers).  
  (Virtual machines have many of the features of minds that have puzzled philosophers.)
– We can explore different architectures for minds, and see which sorts of concepts are appropriate for describing the different sorts of minds  
  (e.g. concepts like ‘perception’, ‘thinking’, ‘emotion’, ‘belief’, ‘pleasure’, ‘consciousness’.)
– We can address the ‘problem of other minds’ (how do we know anything about another mind?) by exploring architectures for agents that need to be able to think about and communicate with other agents.  
  (Different kinds of awareness of other agents in predators, prey, social animals, etc.)
– By attempting to design working models of human minds, and noticing how our programs are inadequate, we discover some unobvious facets of our own minds, and some unobvious requirements (e.g. for perception, learning, reasoning).
Philosophy needs AI and AI needs philosophy

AI needs philosophy to help clarify
- its goals: e.g. what is the study of intelligence? what are intelligent machines?,
- the concepts it uses,
- the kinds of knowledge and ontology a machine needs to interact with us,
- some methodological issues: e.g. how can AI theories be tested? Are the goals of AI achievable?

Philosophy needs AI
- To provide a new context for old philosophical questions.
  E.g. ‘What can be known?’ becomes a more focused question in the context of different specific sorts of machines that can perceive, infer, learn, ...
- To provide a host of new phenomena to investigate, partly to clarify old philosophical concepts and theories. E.g.
  — New kinds of machines: information processing machines
  — New kinds of representation, inference, communication
  — New examples of physical non-physical interaction: virtual machines and computers.
  — New sorts of virtual machine architectures

New examples help refute bad old theories and clarify old concepts.
(See the ‘Philosophical encounter’ in IJCAI 1995 (Minsky, McCarthy and Sloman))
How to think about non-physical levels in reality

At all levels there are objects, properties, relations, structures, mechanisms, states, events, processes and causal interactions.

E.g. poverty can cause crime.

But they are all ultimately implemented in physical systems.

Nobody knows how many levels of virtual machines physicists will eventually discover.

See our IJCAI’01 Philosophy of AI tutorial http://www.cs.bham.ac.uk/~axs/ijcai01/
DIFFERENT VIEWS OF MIND

OLDER APPROACHES:

- A ghost in a machine (dualism)
  - With causal connections both ways: Interactionism
  - With causal connections only one way: Epiphenomenalism
  - With no causal connections: Pre-established harmony
- Mind-brain identity (e.g. the double-aspect theory)
- Behaviourism (mind defined by input-output relations)
- Social/political models of mind
- Mechanical models (e.g. levers, steam engines)
- Electrical models (old telephone exchanges)

PROBLEMS WITH OLDER APPROACHES

- Some lack explanatory power (ghost in the machine)
- Some are circular (Social/Political models of mind)
- Some offer explanations that are too crude to explain fine detail
  and do not generalise (e.g. mechanical and electrical models)

AI provides tools and concepts for developing new rich and precise theories which don’t merely describe some overall structure of mind or mind-body relation, but can show how minds work.
Is there a ghost in the machine?

We now understand that if there is a ghost in the machine it requires sophisticated information-processing capabilities to do what minds do. I.e. there must be a machine in the ghost – an information processing machine.
What AI adds

AI enables philosophy to take account of information processing virtual machines

The Birmingham Cognition and Affect project attempts to develop a new philosophy of mind:

Virtual machine functionalism

See http://www.cs.bham.ac.uk/~axs/lev/

Mental concepts are defined in terms of states and processes in Virtual machines with complex information processing architectures.
Organisms process information

Once upon a time there were only inorganic things: atoms, molecules, rocks, planets, stars, etc. These merely reacted to resultants of all the physical forces acting on them.

Later, there were simple organisms. And then more and more complex organisms.

These organisms had the ability to reproduce. But more interesting was their ability to initiate action, and to select responses, instead of simply being pushed around by resultants.

That achievement required the ability to acquire, process, and use information.

We use “information” in the everyday sense, not the Shannon/Weaver technical sense.
Resist the urge to ask for a definition of “information”

Compare “energy” – the concept has grown much since the time of Newton. Did he understand what energy is?
Instead of defining “information” we need to analyse the following:
– the variety of types of information there are,
– the kinds of forms they can take,
– the means of acquiring information,
– the means of manipulating information,
– the means of storing information,
– the means of communicating information,
– the purposes for which information can be used,
– the variety of ways of using information.

As we learn more about such things, our concept of “information” grows deeper and richer.
Like many deep concepts in science, it is implicitly defined by its role in our theories and our designs for working systems.
Things you can do with information

A partial analysis to illustrate the above:

- You can react immediately (it can trigger immediate action, either external or internal)
- You can do segmenting, clustering labelling of components within a complex information structure (i.e. do parsing)
- You can try to derive new information from it (e.g. what caused this? what else is there? what might happen next? can I benefit from this?)
- You can store it for future use (and possibly modify it later)
- You can consider alternative possibilities, e.g. in planning.
- If you can interpret it as containing instructions, you can obey them, e.g. carrying out a plan.
- You can observe the process of doing all the above and derive new information from it (self-monitoring, meta-management).
- You can communicate it to others (or to yourself later)
- You can check it for consistency, either internal or external

... using different forms of representation for different purposes.
What an organism or machine can do with information depends on its architecture.

Not just its physical architecture – its information processing architecture.

This may be a virtual machine, like

– a chess virtual machine
– a word processor
– a spreadsheet
– an operating system (linux, solaris, windows)
– a compiler
– most of the internet
What is an architecture?

AI used to be mainly about algorithms and representations. Increasingly, during the 1990s and onwards it has been concerned with the study of architectures.

An architecture includes:

– forms of representation,
– algorithms,
– concurrently processing sub-systems,
– connections between them.

Note: Some of the sub-systems may themselves have complex architectures.

We need to understand the space of information processing architectures and the states and processes they can support, including the varieties of types of mental states and processes.

Which architectures can support human-like emotions?
There’s No Unique Correct Architecture

Some tempting wrong ways to think about consciousness:

1. There’s no **continuum** from non-conscious to fully conscious beings

   ![continuum](image)

2. It’s not a **dichotomy** either

   ![dichotomy](image)

Both ‘smooth variation’ and a single discontinuity are poor models.
We need a better view of the space of possibilities

There are many different types of designs, and many ways in which designs can vary.
Some variations are continuous (getting bigger, faster, heavier, etc.).

Some variations are discontinuous:
– duplicating a structure,
– adding a new connection between existing structures,
– replacing a component with another,
– extending a plan.
– adding a new control mechanism

Most biological changes are discontinuous — discontinuities can be big or small.

In particular, evolution produces changes of kind as well as degree.
A simple (insect-like) architecture

A reactive system does not construct descriptions of possible futures evaluate them and then choose one. It simply reacts (internally or externally).

An adaptive system with reactive mechanisms can be a very successful biological machine. Some purely reactive species also have a social architecture, e.g. ants, termites, and other insects.
Features of reactive organisms

The main feature of reactive systems is that they lack the ability to represent and reason about non-existent phenomena (e.g. future possible actions), the core ability of deliberative systems, explained below.

Reactive systems need not be “stateless”: some internal reactions can change internal states, and that can influence future reactions.

In particular, reactive systems may be adaptive: e.g. trainable neural nets, which adapt as a result of positive or negative reinforcement.

Some reactions will produce external behaviour. Others will merely produce internal changes.

Internal reactions may form loops.

An interesting special case are teleo-reactive systems, described by Nilsson (http://robotics.stanford.edu/)

In principle a reactive system can produce any external behaviour that more sophisticated systems can produce: but possibly requiring a larger memory for pre-stored reactive behaviours than could fit into the whole universe. Evolution seems to have discovered the advantages of deliberative capabilities.
Is a fly conscious of the hand swooping down to kill it? Insects perceive things in their environment, and behave accordingly. However, it is not clear whether their perceptual mechanisms produce information states between perception and action usable in different ways in combination with different sorts of information. (Compare ways of using information that a table is in the room.)

Rather, it seems that their sensory inputs directly drive action-control signals, though possibly after transformations which may reduce dimensionality, as in simple feed-forward neural nets.

There may be exceptions: e.g. bees get information which can be used either to control their own behaviour or to generate “messages” that influence the behaviour of others.

Typically a purely reactive system does not use information with the same type of flexibility as a deliberative system which can consider non-existent possibilities.

They also lack self-awareness, self-categorising abilities. A fly that sees an approaching hand probably does not know that it sees — it lacks meta-management mechanisms, described later.
Give REACTIVE DEMO

Sheepdog
‘Emotional’ agents
Sometimes the ability to plan is useful

Deliberative mechanisms provide the ability to represent possibilities (e.g. possible actions, possible explanations for what is perceived). Much, but not all, early symbolic AI was concerned with deliberative systems (planners, problem-solvers, parsers, theorem-provers).
Give DELIBERATIVE DEMO

SHRDLU (pop11 gblocks)
Deliberative mechanisms

These differ in various ways:

- the forms of representations (often data-structures in virtual machines)
- the variety of forms available (e.g. logical, pictorial, activation vectors)
- the algorithms/mechanisms available for manipulating representations
- the number of possibilities that can be represented simultaneously
- the depth of ‘look-ahead’ in planning
- the ability to represent future, past, or remote present objects or events
- the ability to represent possible actions of other agents
- the ability to represent mental states of others (linked to meta-management, below).
- the ability to represent abstract entities (numbers, rules, proofs)
- the ability to learn, in various ways

Some deliberative capabilities require the ability to learn new abstract associations, e.g. between situations and possible actions, between actions and possible effects
Evolutionary pressures on perceptual and action mechanisms for deliberative agents

New levels of perceptual abstraction (e.g. perceiving object types, abstract affordances), and support for high-level motor commands (e.g. “walk to tree”, “grasp berry”) might evolve to meet deliberative needs – hence taller perception and action towers.
Multi-window perception and action

If multiple levels and types of perceptual processing go on in parallel, we can talk about

“multi-window perception”,

as opposed to

“peephole” perception.

Likewise, in an architecture there can be

multi-window action

or merely

peephole action.
Did Good Old Fashioned AI (GOFAI) fail?

It is often claimed that symbolic AI and the work on deliberative systems failed in the 1970s and 1980s and therefore a new approach to AI was needed. New approaches (some defended by philosophers) included use of neural nets, use of reactive systems, use of subsumption architectures (Rodney Brooks), use of evolutionary methods (genetic algorithms, genetic programming) and use of dynamical systems (using equations borrowed from physics and control engineering).

The critics missed the point that many of the AI systems of the 1970s and 1980s were disappointing partly because they used very small and very slow computers (e.g. 1MByte was a huge amount of memory in 1980), partly because they did not have enough knowledge about the world, and partly because the architecture lacked self-monitoring capabilities: meta-management.

The new emphasis on architectures helps us think more clearly about combining components required to match human capabilities.
The pressure towards self-knowledge, self-evaluation and self-control

A deliberative system can easily get stuck in loops or repeat the same unsuccessful attempt to solve a sub-problem. One way to prevent this is to have a parallel sub-system monitoring and evaluating the deliberative processes. If it detects something bad happening, then it may be able to interrupt and re-direct the processing.

(Compare Minsky on “B brains” and “C brains” in *Society of Mind*)

We call this meta-management. It seems to be rare in biological organisms and probably evolved very late.

As with deliberative and reactive mechanisms, there are many forms of meta-management.

Conjecture: the representational capabilities that evolved for dealing with self-categorisation can also be used for other-categorisation, and vice-versa. Perceptual mechanisms may have evolved recently to use these representational capabilities in percepts.

Example: seeing someone else as happy, or angry.
Later, meta-management (reflection) evolved

A conjectured generalisation of homeostasis.

Self monitoring, can include categorisation, evaluation, and (partial) control of internal processes. Not just measurement.

The richest versions of this evolved very recently, and may be restricted to humans.

Research on ‘reflective’ AI systems is in progress.

Absence of meta-management can lead to stupid behaviour in AI systems, and in brain-damaged humans.

Further steps to a human-like architecture

CONJECTURE:

Central meta-management led to opportunities for evolution of

− additional layers in ‘multi-window perceptual systems’
  and
− additional layers in ‘multi-window action systems’,

Examples: social perception (seeing someone as sad or happy or puzzled), and stylised social action, e.g. courtly bows, social modulation of speech production.

Additional requirements led to further complexity in the architecture, e.g.

− ‘interrupt filters’ for resource-limited attention mechanisms,
− more or less global ‘alarm mechanisms’ for dealing with important and urgent problems and opportunities,
− socially influenced store of personalities/personae

All shown in the next slide, with extended layers of perception and action.
More layers of abstraction in perception and action, and global alarm mechanisms

This conjectured architecture (H-Cogaff) could be included in robots (in the distant future). Arrows represent information flow (including control signals).

If meta-management processes have access to intermediate perceptual databases, then this can produce self-monitoring of sensory contents, leading robot philosophers with this architecture to discover “the problem(s) of Qualia?”

‘Alarm’ mechanisms can achieve rapid global re-organisation.
Some Implications

Within this framework we can explain (or predict) many phenomena, some of them part of everyday experience and some discovered by scientists:

- Several varieties of **emotions**: at least three distinct types related to the three layers: **primary** (exclusively reactive), **secondary** (partly deliberative) and **tertiary** emotions (including disruption of meta-management) – some shared with other animals, some unique to humans. (For more on this see Cogaff Project papers)

- Discovery of **different visual pathways**, since there are many routes for visual information to be used.
  (See talk 8 in http://www.cs.bham.ac.uk/~axs/misc/talks/)

- Many possible **types of brain damage** and their effects, e.g. frontal-lobe damage interfering with meta-management (Damasio).

- **Blindsight** (damage to some meta-management access routes prevents self-knowledge about intact (reactive?) visual processes.)

This helps to enrich the analyses of concepts produced by philosophers sitting in their arm chairs: for it is very hard to dream up all these examples of kinds of architectures, states, processes if you merely use your own imagination.
Implications continued ....

- Many varieties of learning and development
  (E.g. “skill compilation” when repeated actions at deliberative levels train reactive systems to produce fast fluent actions, and action sequences. Needs spare capacity in reactive mechanisms, (e.g. the cerebellum?). We can also analyse development of the architecture in infancy, including development of personality as the architecture grows.)

- Conjecture: mathematical development depends on development of meta-management – the ability to attend to and reflect on thought processes and their structure, e.g. noticing features of your own counting operations, or features of your visual processes.

- Further work may help us understand some of the evolutionary trade-offs in developing these systems.
  (Deliberative and meta-management mechanisms can be very expensive, and require a food pyramid to support them.)

- Discovery by philosophers of sensory ‘qualia’. We can see how philosophical thoughts (and confusions) about consciousness are inevitable in intelligent systems with partial self-knowledge.

For more see papers here: http://www.cs.bham.ac.uk/research/cogaff/
How to explain qualia

Philosophers (and others) contemplating the content of their own experience tend to conclude that there is a very special type of entity to which we have special access only from inside qualia (singular is ‘quale’). This generates apparently endless debates.

For more on this see talk 12 on consciousness here http://www.cs.bham.ac.uk/~axs/misc/talks/

We don’t explain qualia by saying what they are.

Instead we explain the phenomena that generate philosophical thinking of the sort found in discussions of qualia.

It is a consequence of having the ability to attend to aspects of internal information processing (internal self-awareness), and then trying to express the results of such attention.

That possibility is inherent in any system that has the sort of architecture we call H-Cogaff, though different versions will be present in different architectures, e.g. depending on the forms of representation and modes of monitoring available to meta-management.

Robots with that architecture may also ‘discover’ qualia.
Families of architecture-based mental concepts

For each architecture we can specify a family of concepts of types of virtual machine information processing states, processes and capabilities supported by the architecture.

Theories of the architecture of matter refined and extended our concepts of kinds of stuff (periodic table of elements, and varieties of chemical compounds) and of physical and chemical processes.

Likewise, architecture-based mental concepts can extend and refine our semantically indeterminate pre-theoretical concepts, leading to much clearer concepts related to the mechanisms that can produce different sorts of mental states and processes.

Philosophy will never be the same again.
New questions supplant old ones

We can expect to replace old unanswerable questions.

Is a fly conscious? Can a foetus feel pain?

is replaced by new EMPIRICAL questions, e.g.

Which of the 37 varieties of consciousness does a fly have, if any?

Which types of pain can occur in an unborn foetus aged N months and in which sense of ‘being aware’ can it be aware of them, if any?

Of course, this may also generate new ethical questions, about the rights of robots and our duties towards them.

And that will feed new problems into moral philosophy.
The causation problem: Epiphenomenalism

A problem not discussed here is how it is possible for events in virtual machines to have causal powers.

It is sometimes argued that since (by hypothesis) virtual machines are fully implemented in physical machines, the only causes really operating are the physical ones.

This leads to the conclusion that virtual machines and their contents are “epiphenomenal”, i.e. lacking causal powers.

If correct that would imply that if mental phenomena are all states, processes or events in virtual information processing machines, then mental phenomena (e.g. desires, decisions) have no causal powers.

A similar argument would refute many assumptions of everyday life, e.g. ignorance can cause poverty, poverty can cause crime, etc.

Dealing with this issue requires a deep analysis of the notion of ‘cause’, probably the hardest unsolved problem in philosophy.

A sketch of an answer is offered in this Philosophy of AI tutorial presentation:  http://www.cs.bham.ac.uk/~axs/ijcai01

See also the talk on supervenience and implementation in http://www.cs.bham.ac.uk/~axs/misc/talks/
A problem

Too many philosophers are ignorant of AI and Computer Science, and have over-simplified ideas about what they do. So they produce false generalisations about computers can or cannot do.

Too many AI researchers are ignorant of philosophy and do not have good philosophical analytical skills. So they do not detect muddle and confusion in their own concepts.

MAYBE YOU CAN STUDY BOTH AND HELP TO IMPROVE BOTH DISCIPLINES IN THE FUTURE?

(Similar comments can be made about psychology and AI.)