The Philosophy and Ethics of Artificial Intelligence

Dr. Sanjay Modgil
N7.08 Bush House and Finchley, North London

Philosophy and AI

- What is philosophy
- Why is philosophy relevant to AI
- How philosophical issues impact on the ethics of AI
- Overview of Module structure and Assessment

What is Philosophy ?

- “Philosophy" literally means, "love of wisdom."
- In a broad sense, philosophy is an activity people undertake when they seek to understand fundamental truths about themselves, the world in which they live, and their relationships to the world and to each other.
- As an academic discipline philosophy is much the same. Those who study philosophy are engaged in asking, answering, and arguing for their answers to life’s most basic questions. Why are we here ? Why is there something rather than nothing ? What is knowledge ? Why are we, and what is, consciousness ? What is the good life ? What are the morally right things to do ? …..
- Philosophy is traditionally divided into major areas of study

What is Philosophy ?

Metaphysics is the study of the nature of reality, of what exists in the world, what it is like, and how it is ordered. Typical metaphysical questions are:

- Is there a God?
- What is truth?
- What is a person? What makes a person the same through time? What is the self ?
- Is the world strictly composed of matter?
- Do people have minds? If so, how is the mind related to the body?
- Do people have free will ?

Epistemology is the study of knowledge. It is primarily concerned with what we can know about the world and how we can know it. Typical questions are:

- What is knowledge?
- Do we know anything at all?
- How do we know what we know is true ?
- Can we be justified in claiming to know certain things?

What is Philosophy ?

Ethics concerns what we ought to do and what it would be best to do. In struggling with this issue, larger questions about what is good and right arise. So, the ethicist attempts to answer such questions as

- What is good? What makes actions or people good?
- What is right? What makes actions right?
- Is morality objective or subjective?
- How should I treat others?

Logic An important aspect of the study of philosophy is the arguments or reasons given for answers to questions. Philosophers employ logic to study the nature and structure of arguments. Logicians ask such questions as:

- What constitutes "good" or "bad" reasoning?
- How do we determine whether a given piece of reasoning is good or bad?

Fundamental philosophical questions arise in almost every discipline. And so we have philosophy of law, philosophy of mind, philosophy of religion, philosophy of science, political philosophy, philosophy and technology …

What is the impact of Philosophy ?

Example 1 ("I think therefore I am")

In 1637 the famous philosopher and mathematician Rene Descartes wrote "I think therefore I am" to highlight that the very act of thinking establishes an absolute proof of the existence of a self (the I) that is doing the thinking. Whereas everything else – the external world we perceive – could be an illusion conjured up by a demon (we're in the Matrix !) to deceive us.

The only fact that we can be absolutely certain of is the existence of the self. But we cannot be absolutely certain about anything else, including the physical world the body (we could be a brain wired up to machine). Therefore the self/*mind* must be something that is fundamentally different from the physical world.

This separation of mind and body (until relatively recently) permeated ***Western*** thought. So for example, in western medicine, the idea that one's mental processes (e.g., positive thinking) could have an effect on one's physical state would have been dismissed. But not now !

What is the impact of Philosophy ?

Example 2 ("Free will, the self, responsibility and justice")

- Physics ultimately describes the nature and workings of the world including human behaviour. All physical processes are *deterministic* – A is caused by B is caused by C and so on …

- Free will is therefore an illusion – the illusion that we have a 'genuine' choice, rather that a choice determined by underlying neural processes in the brain, influenced by inputs from the environment

- So when we hold a criminal responsible for a crime, there is no *self* making a 'free' decision that we hold responsible. The committal of the crime was a consequence of a deterministic series of causes and effects

- Therefore penalties for a crime should not be based on a desire to punish a self that is morally responsible. Rather, the penalty should be such that: 1) it is incorporated into the chain of reasoning responsible for future decisions, such that decisions will not be made that result in crimes (*rehabilitation*), and 2) acts as a *deterrent* to others, and 3) protects society from the criminal and the criminal from society

What is the impact of Philosophy ?

- So when we anticipate future AIs making decisions in the future, and when things go wrong, and harms are caused, to whom or what, and how, should we assign moral responsibility, and what actions should we take to prevent harms in the future ?

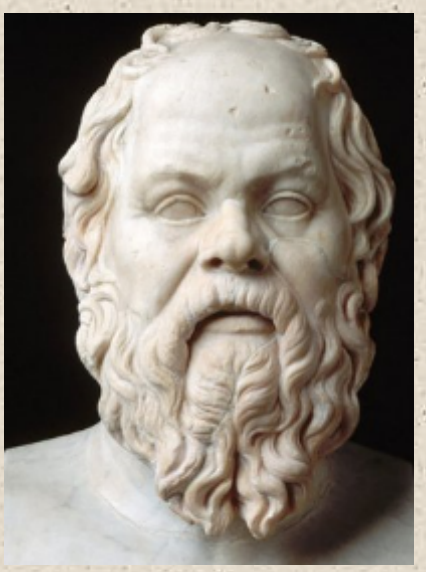
- For example, autonomous cars, robots in the battlefield

What is the impact of Philosophy ?

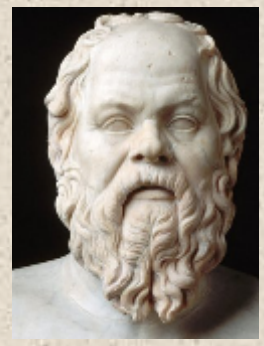
Example 3 (“Logical Argument and Ethics”)

- An example of how you can use logic in argument and debate

- Socrates used a method of questioning to expose weaknesses (contradictions) in his adversaries’ arguments



The abortion debate : Socrates v Sara



Why ?

So ?

Abortion is wrong

Human life is sacred and a foetus is a human

If a life is sacred it should not be terminated

The abortion debate : Socrates v Sara

- By questioning Sara, Socrates has revealed Sara's line of reasoning / argument

If X is a human then X's life is sacred

A foetus is a human

A foetus's life is sacred

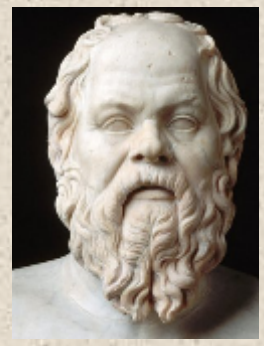
If X's life is sacred then X should not be terminated

A foetus's life is sacred

A foetus should not be terminated

t

The abortion debate : Socrates v Sara



Why ?

So ?

So you should be against the death penalty for murderers ?

Abortion is wrong

Human life is sacred and a foetus is a human

If a life is sacred it should not be terminated

Errr, no ! I'm in favour

The abortion debate : Socrates v Sara

- Socrates has revealed a contradiction in, and so weakened, Sara's argument against abortion, since given her argument against abortion, she should be ***against*** the death penalty, but she is in favour !

If X is a human then X's life is sacred
A murderer is a human

A murderer's life is sacred

If X's life is sacred then X should not be terminated
A murderer's life is sacred

A murderer should not be terminated

t

The abortion debate : Socrates v Sara

- Knowledge often evolves through refutation (e.g., scientific knowledge) so that for example seeing a bird (penguin) flapping its wings but hopelessly failing to fly, we may change the rule

If X is a bird then X flies

to

If X is a bird and X is not a penguin then X flies

so that there is no longer a contradiction

- How can Sara change the knowledge she uses in her debate with Socrates so that her being in favour of the death penalty does not contradict her being against abortion?

Refining knowledge through refutation

If X is a human and **X does not take the life of another human** then X's life is sacred

t

Refining knowledge through refutation

If X is a human and X does not take the life of another human then X's life is sacred

A foetus is a human and a foetus does not take the life of another human

---

A foetus's life is sacred

If X's life is sacred then X should not be terminated

A foetus's life is sacred

A foetus should not be terminated

t

Refining knowledge through refutation

If X is a human and X does not take the life of another human then X's life is sacred

A murderer is a human BUT a murderer **does** take the life of another human

So Sara is **not** forced to conclude from the same line of reasoning for foetuses, that a murderer should not be terminated

How do you think Socrates might continue in this debate ?

Overview of Course

- Introduction
- Intelligence
- Consciousness
- Reasoning and Communication
- Ethics and Morality
- Algorithms
- For good or bad ? AI in Medicine, AI in War
- Superintelligence and the Value Loading Problem
- AI and Human Society
- Review/Revision

Intelligence

- What is Intelligence ?
- What is *Artificial* Intelligence ?
- *Narrow*, *General* and *Super* Intelligence
- How can we tell that a machine is intelligent ?
- Arguments against the possibility of Artificial Intelligence

Consciousness

- What is Consciousness ?
- Problems of Consciousness
- Theories of Consciousness
- Could an AI be conscious ?
- Ethical Implications

Reasoning and Communication

- The major AI paradigms (Symbolic v Non-Symbolic (ML))
- Limitations of paradigms
- Non-monotonic Reasoning and Argumentation
- Communication

Ethics and Morality

- Ethical Challenges for Artificial Intelligence
- Philosophical Accounts of Ethics and Morality
- Implementing Moral Agents
- Utilitarianism

Algorithms

- Accountability, Responsibility and Justice
- Algorithmic Bias
- Towards a Professional Code of Ethics

For good or bad ? AI in Medicine, AI in War

- Lethal Autonomous Weapons: Arguments for and Against Banning them
- AI in Medicine. Examples and Ethical Issues

Superintelligence and the Value Loading Problem

- The Path from Artificial General Intelligence to Superintelligence
- What Superintelligence Machines can and might do
- Why should we be worried about the dangers of Superintelligence
- The Value Loading Problem and potential solutions

AI and Human Society

- How will humanity be changed by AI, AGI and Superintelligence ?
- Belief Bubbles, Echo Chambers and Surveillance Capitalism
- Thinking out of the box: How AGI and Superintelligence might help contribute to moral/ethical progress

Is this Module for You ?

- Prerequisites for this course are a **basic** background in AI and logic (obtained to a significant extent by your taking other modules)
- An interest in critical thinking, discussion debate and argument, philosophy
- An interest in how AI will develop, and how should be developed in the future, in particular to ensure that AI is used for our benefit and not for harm
- There will be limited amount of technical content; most of the content will be conceptual, discursive, **and make relatively sophisticated use of language and argument**

Module Structure: Lectures and Tutorials

- Each week I will make available on the KEATs module page, a video and pdf of the lectures and you will have a week to go through the material.

- One week after the lecture material is made available on KEATs, there will be a large group tutorial in which you can raise any questions with me – these are opportunities for further discussion

- Each week there will also be small group tutorials led by TAs (Teaching Assistants) who will present additional material related to the topics covered in my lectures

Weekly Lecture Topics

- **Introductory Lecture**
- **Intelligence**
- **Consciousness**
- **Reasoning and Communication**
- **Ethics and Morality**
- **Algorithms**
- **AI in Medicine and Law**
- **Superintelligence and the Value Loading Problem**
- **AI and Human Society**
- **Revision Lecture**

Assessment

- Online Exam in January: two hour multiple choice question exam. During the semester I will set example MCQ questions to give you a feel for the kinds of questions you will have in your exam.

- **You will only be examined on material in the lecture slides**. Tutorials will deepen and consolidate your understanding and critical reasoning skills.

- Pdfs of lecture slides and all tutorial materials will be made available on KEATs.

Sources

The following books and online resources are a selection of sources for this course – you can consult these for your own interest and to deepen your understanding, but do not need to consult these in order to pass the exam.

- *Life 3.0*. Max Tegmark
- Stanford Encyclopedia of Philosophy (https://plato.stanford.edu/index.html)
- *Superintelligence: Paths, Dangers, Strategies.* Nick Bostrom.
- *Moral Machines: Teaching Robots Right from Wrong.* Wendell Wallach and Colin Allen
- *Towards a Code of Ethics for Artificial Intelligence*. Paula Boddington
- *Moral Tribes*. Joshua Greene

Help

- If you have any questions regarding the content of the **lectures first use the Discussion & Advice Forum on the module's KEATs page**, in which other students can reply with answers, and if the TAs (Yani and Alex) or I feel that we need to clarify answers provide by other students, we will jump in

- Each Large Group Tutorial when you can ask me questions about the previous week's lecture materials.

- You can also ask me questions during my office hours (see details on the KEATs module page)

- **Do not use Teams to contact lecturers or TAs**

Expectations of behaviour

Staff and students are expected to behave respectfully to one another – during lectures, outside of lectures and when communicating online or through email.

We won't tolerate inappropriate or demeaning comments related to gender, gender identity and expression, sexual orientation, disability, physical appearance, race, religion, age, or any other personal characteristic.

If you witness or experience any behaviour you are concerned about, please speak to someone about it. This could be one of your lecturers, your personal tutor, a programme administrator, the diversity & inclusion co-chairs (Alfie Abdul-Rahman & Kathleen Steinhöfel) at informatics-diversity@kcl.ac.uk, a trained harassment advisor, or any member of staff you feel comfortable talking about it to. More info at https://www.kcl.ac.uk/hr/diversity/dignity-at-kings