Lingnan University
Department of Philosophy

Course Title	:	A Philosophical Introduction to Artificial Intelligence
Course Code	:	PHI2118
Recommended Study Year	:	2nd and 3rd years
No. of Credits/Term	:	3
Mode of Tuition	:	Lecture and Tutorial
Class Contact Hours	:	Two hours lecture per week One hour tutorial per week
Category in Major Programme	:	Programme Elective
Prerequisite(s)	:	N/A
Co-requisite(s)	:	N/A
Exclusion(s)	:	PHI4375 Philosophy of Artificial Intelligence
Exemption Requirement(s)	:	N/A

Brief Course Description

In the past fifty years, research in Artificial Intelligence (AI), which involves the attempt to build intelligent computers, has led to the production of a multitude of autonomous machines, many of which are now in common use; for example, machines that autonomously perform medical diagnoses, artificial pets and companions in health- and eldercare, self-driving cars, face recognition programs and war robots. This course provides an introduction to the technical and philosophical issues concerning the creation and nature of these artificially intelligent machines. The course also offers a wide-ranging introduction to some of the implications of AI for human life, society and culture.

Aims

The course aims at:

1. Providing students with an understanding of the fundamental principles of modern computing approaches to Artificial Intelligence (expert systems, genetic algorithms, neural nets), insofar as they are necessary for the understanding of the philosophical problems of AI.
2. Presenting some of the central, currently discussed problems and key concepts in the field of Philosophy of Artificial Intelligence: robot ethics, the responsibility for machine actions, the topic of emotional attachment between people and artefacts, as well as the discussion about the conditions of the personhood of humans and non-humans.
3. Enabling students to question their own conceptions of machine intelligence and personhood in an informed, argumentative way.
4. Providing students with the means to appreciate the connection of the questions posed by new AI developments with similar questions in Bioethics, philosophy of mind and normative ethics (responsibility and liability).

Learning Outcomes

On completion of the course, students will be able to:

- (LO1) Identify and interpret basic technical notions employed in contemporary Artificial Intelligence, and what problems each one of them poses.
- (LO2) Evaluate critically the main arguments for and against the notion of 'machine intelligence'.
- (LO3) Reflect on the relationship between modern technology and the cultural assumptions that underlie it, especially in the areas of machine responsibility and personhood.
- (LO4) Assess the main ethical problems posed by the use of autonomous machines.

Indicative Content

Part I: Principles and architecture of AI systems

1. Finite State Machines and their use in AI
2. Programming symbolic AI in Prolog
3. Expert systems and their critics
4. CYC and common-sense computing

5. Nouvelle and subsymbolic AI: Braitenberg vehicles and subsumption architecture
6. Neural networks I & II, reinforcement learning, genetic computing
7. Machine learning terminology and main directions
8. Connectomes, hybrots and cyborgs

Part II: Philosophical problems of AI

9. The Physical Symbol System Hypothesis
10. Computationalism, The Turing Test and its critics
11. Can and should machines have a mind?
12. The Extended Mind: the environment as mind

Part III: Applications of AI

13. Responsibility and liability of machines
14. Emotional machines and loving machines
15. The ethics of self-driving cars
16. AI and jobs: Employment and universal basic income
17. AI and democracy
18. Law enforcement, military and killer robots

Teaching Method
Lecture, tutorial and experiential activities (demonstration of Artificial Intelligence systems, chatbots, neural nets). Screening of sections of relevant films. Research for related news topics and analysis of the problems posed in them.

Measurement of Learning Outcomes

1. Continuous assessment, in the form of in-class quizzes and discussions, to measure LO1, LO2, LO3. Students will have to demonstrate having the capacity to reflect deeply and in an informed manner on the issues related to the session's topic.
2. Students will write an individual term paper. They are expected to be able to integrate what they have learned in class with their own research in news and scholarly publications in order to apprehend concrete situations. To measure LO1, LO2, LO3, LO4.
3. A final examination which will assess students' understanding of the classical debates of the Philosophy of Artificial Intelligence, the problems of machine responsibility and personhood, and the main ethical points regarding the use of autonomous robots and bionic organisms (LO2, LO3, LO4)

Assessment
Continuous Assessment (70%) – in-class quizzes 10%, in-class discussion 20%, term paper 40%.
Final examination (30%)

The final examination will consist of short answer questions and essay questions.

Required Readings
Selections from
[Collections]
Born, R. (ed.) Artificial Intelligence. The Case Against. London/New York: Routledge. 1988.
Graubard, S.R. (ed.) The Artificial Intelligence Debate. False Starts, Real Foundations. Cambridge, Mass.: MIT Press. 1988.

Supplementary Readings
[Collections]
Gill, K.S. (ed.) Artificial Intelligence for Society. Chichester etc: John Wiley. 1986.
Torrance, S.B. (ed.) The Mind and the Machine. Philosophical Aspects of Artificial Intelligence. New York etc:

Ellis Horwood. 1984.

[General]
Anderson, D. Artificial Intelligence and Intelligent Systems. The Implications. Chichester etc: Ellis Horwood. 1989.
Boden, Margaret A. Artificial intelligence: a very short introduction. Oxford University Press. 2018
Levy, S. Artificial Life. The Quest for a New Creation. London: Jonathan Cape. 1992.
Mitchell, Melanie. Artificial intelligence: A guide for thinking humans. Penguin UK, 2019.
Russell, S. and Norvig P. Artificial Intelligence. A Modern Approach. Pearson Series in Artificial Intelligence. 4th edition. Pearson. 2020.
Sharples, M. and Hogg, D. and Hutchinson, C. et al. Computers and Thought. A Practical Introduction to Artificial Intelligence. Cambridge, Mass.: MIT Press. 1989.

[Psychology]
Boden, M.A. Artificial Intelligence and Natural Man. 2nd expanded ed. New York: Basic Books. 1987.

[Turing Test]
Saygin, A.P., Cicekli, I. and Akman, V. "Turing Test: 50 Years Later". Minds and Machines 10, No. 4 (2008): 463-518.
Shieber, S. M. (1994). Lessons from a restricted Turing test. arXiv preprint cmp-lg/9404002.
Whitby, B. (2000). AI's biggest blind alley?. Artificial Intelligence: Critical Concepts, 4, 195.

[Chinese Room]
Mooney III, V. J. (1997). Searle's Chinese room and its aftermath. Center for the Study of Language and Information Report No. CSLI, 97, 202.
Preston, J. and Bishop, M. (eds.) Views into the Chinese Room. New Essays on Searle and Artificial Intelligence. Oxford: Clarendon. 2002.

[Various]
Churchland, P.S. Neurophilosophy. Toward a Unified Science of the Mind/Brain. Cambridge, Mass.: MIT Press. 1986.
Dennett, D.C. Brainstorms. Philosophical Essays on Mind and Psychology. Cambridge, Mass.: Bradford/MIT Press. 1978.
Dennett, D.C. The Intentional Stance. Cambridge Mass./London: MIT Press, 1987
Minsky, M. Why People Think Computers Can't. AI Magazine, 3 No. 4 (1982).

[AI techniques]
Holland, J. H. (1992). Genetic algorithms. Scientific American, 267(1), 66-73.
Lenat, D. CYC: Toward Programs With Common Sense. Communications of the ACM 33, No. 8 (1990): 30 ff.
Minsky, M.L. The Society of Mind. London: Heinemann. 1987.
Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36-45.

[AI critique]
Clark, Andy. Mindware: An introduction to the philosophy of cognitive science. 2nd ed. Oxford University Press, 2014.
Dreyfus, H.L. What Computers Still Can't Do. A Critique of Artificial Reason. Cambridge, Mass.: MIT Press. 1992.
Dreyfus, H. L. (2002). Intelligence without representation. Phenomenology and the cognitive sciences, 1(4), 367-383.
Dreyfus, H. L., & Dreyfus, S. E. (1986). From Socrates to expert systems: The limits of calculative rationality. In Philosophy and technology II (pp. 111-130). Springer, Dordrecht.
McClintock, A. The Convergence of Machine and Human Nature. A Critique of the Computer Metaphor of Mind and Artificial Intelligence. Aldershot: Avebury. 1995.

[Responsibility and Personhood]
Dworkin, G. Intention, Foreseeability, and Responsibility. In Schoeman, F. (ed.): Responsibility, Character, and the Emotions. New Essays in Moral Psychology. Cambridge University Press, 1987: 338–354
Fischer, J.M. and Ravizza, M.S. Responsibility and Control. A Theory of Moral Responsibility. Cambridge University Press. 1998
Wolf, S. Sanity and the Metaphysics of Responsibility. In Schoeman, F. (ed.): Responsibility, Character, and the

Emotions. New Essays in Moral Psychology. Cambridge University Press, 1987: 46–62

Important Notes

(1) Students are expected to spend a total of 9 hours (i.e. 3 hours of class contact and 6 hours of personal study) per week to achieve the course learning outcomes.
(2) Students shall be aware of the University regulations about dishonest practice in course work, tests and examinations, and the possible consequences as stipulated in the Regulations Governing University Examinations. In particular, plagiarism, being a kind of dishonest practice, is "the presentation of another person's work without proper acknowledgement of the source, including exact phrases, or summarised ideas, or even footnotes/citations, whether protected by copyright or not, as the student's own work". Students are required to strictly follow university regulations governing academic integrity and honesty.
(3) Students are required to submit writing assignment(s) using Turnitin.
(4) To enhance students' understanding of plagiarism, a mini-course "Online Tutorial on Plagiarism Awareness" is available on https://pla.ln.edu.hk/