Lingnan University
Department of Philosophy

Course Title	:	Philosophy of Artificial Intelligence
Course Code	:	PHI4375
Recommended Study Year	:	3rd Year
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 - Philosophy of Natural and Human Sciences Profile
Prerequisite(s)	:	N/A
Co-requisite(s)	:	N/A
Exclusion(s)	:	N/A
Exemption Requirement(s)	:	N/A

Brief Course Description

The dream of creating intelligent artifacts is as old as mankind. In the past fifty years, Artificial Intelligence as the attempt to build "intelligent" computers, has had a history rich with hopes and disappointments, and has nonetheless produced a multitude of autonomous machines, which are now in common use. Adaptive and autonomous machines stop being mere tools and become agents in social processes, from machines which autonomously diagnose cancer in the Chinese countryside, to artificial pets like the AIBO or the Tamagotchi, which become partners in emotional exchanges with their owners, to face recognition programs and war robots. We will follow the traces of the attempt to build the "artificial man" through history, from Daedalus' talking statues in Ancient Greece, to the Turing Test for machine intelligence. Lately, the question of the responsibility for machine actions has attracted some attention, and in this course we will present the main arguments in this discussion. After presenting the fundamentals of the technology which underlies modern adaptive software engineering, we will present traditional as well as contemporary arguments in the field of the philosophy of Artificial Intelligence, and discuss the possible conditions for the personhood of machines, as well as the moral implications of autonomously acting robots and bionic organisms.

Aims

The course aims at:

- Providing students with an understanding of the fundamental principles of modern computing approaches to Artificial Intelligence (expert systems, genetic algorithms, neural nets), and their unique qualities and limitations, insofar as they are necessary for the understanding of the philosophical problems of AI.
- Introducing students to the history of Artificial Intelligence, from mythical antiquity to the present.
- Presenting some of the classical discussions and arguments related to Artificial Intelligence.
- Presenting some of the central, currently discussed problems and key concepts in the field of Philosophy of Artificial Intelligence: the concept of intentionality, the distinction between "real" and "simulated", the question of responsibility for machine actions, the topic of emotional attachment between man and artifact, as well as the discussion about the conditions for the personhood of humans and non-humans.
- Enabling students to question their own conceptions of machine intelligence and personhood in an informed, argumentative way.
- Providing students with a broader view of the subject, so that they perceive the connection of

the questions posed by new AI developments with similar questions in Bioethics, in theoretical philosophy (philosophy of mind, personhood), and in ethics (responsibility and liability).

Learning Outcomes

At the end of the course, the students are expected to be able to make informed judgments about the main philosophical consequences of Artificial Intelligence technology.

1. They will understand the different methods and tools employed, and what problems each one of them poses.
2. They will be able to evaluate critically the main arguments for and against the notion of "machine intelligence".
3. They will be able to reflect about the relationship between modern technology and the cultural assumptions that underlie it, especially in the areas of machine responsibility and personhood.
4. They will be able to assess the main ethical problems posed by the use of autonomous machines.

Indicative Content

- The artificial man in myth, art and history.
- Methods of Artificial Intelligence: expert systems, common-sense computing (CYC), genetic programming, neural nets. Embodied AI (mobile agemnts) and Brooks' subsumption architecture.
- Paradigms of natural language processing: Eliza, SHRDLU and Alice. The Turing Test for machine intelligence. The Loebner Test and the critique of Turing-type tests.
- Bionic organisms (Steve Potter). Bionic creatures and the problems of the definition of life.
- Affective Artificial Intelligence: Lyotard, Tamagotchi, reactive attitudes of humans towards machines. "Having" emotions.
- Real and simulated: Is a simulated emotion a real emotion? Is simulated intelligence real intelligence?
- The early AI debate: Marvin Minsky's Society of Mind and Hubert Dreyfus' critique of calculative rationality.
- What is understanding? Searle's Chinese room and its implications.
- Intentional systems: Dennett's approach and its limitations.
- Conditions of responsibility: Conditions for the loss of human responsibility, conditions for reduced, nonhuman and machine responsibility. Dennett, Fisher/Ravizza, Dworkin, S. Wolf. Corporate compared to machine responsibility.
- Conditions for the personhood of humans and nonhumans. Personhood of functionally reduced humans, of animals, of fictional and religious figures, of the dead. The reduced personhood of humans in working environments. The relativity of personhood concepts.
- Ethical consequences of the deployment of AI technologies: AI for the care of the elderly, war robots, face recognition in surveillance cameras, moral status of bionic creatures.

Teaching Method

Lecture, tutorial and experiental 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

- Students will discuss on assigned topics in the tutorials. They are expected to be able to reflect deeply and in an informed manner on the issues related to the session's topic. (LO#1, #2)
- Students will write a 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. (LO#2, #3, #4)

- The examination 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. (LO#1, #2, #3, #4)

Assessment
Continuous assessment, including presentations: 30%. Term paper: 40%. Final examination: 30%.

Required Readings
Selections from
[Collections]
Carter, M. Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence. Edinburgh University Press, 2007.
Feigenbaum, E.A. and Feldman, J. (eds.). Computers & Thought. Menlo Park etc: AAAI Press, MIT Press. 1995.
Lycan, W.G. and Prinz, J.J. (eds). Mind and cognition: an anthology. Malden, MA: Blackwell Pub. Ltd. 2008.

Supplementary Readings
[Collections]
Boden, M.A. (ed.) The Philosophy of Artificial Intelligence (Oxford Readings in Philosophy) Oxford University Press. 1990.
Boden, M.A. (ed.) The Philosophy of Artificial Life (Oxford Readings in Philosophy) Oxford University Press. 1996.
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.
Levy, S., Artificial Life. The Quest for a New Creation. London: Jonathan Cape. 1992.
Scientific American, Understanding Artificial Intelligence (Science Made Accessible) Grand Central Publishing, 2002.
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]
Loebner, H. In Response. (undated) http://www.loebner.net/Prizef/In-response.html.
Mauldin, M.M. Chatterbots, Tinymuds, And The Turing Test: Entering The Loebner Prize Competition. Paper presented at AAAI-94, 1994. http://www.lazytd.com/lti/pub/aaai94.html
Saygin, A.P., Cicekli, I. and Akman, V. "Turing Test: 50 Years Later". Minds and Machines 10, No. 4 (2008): 463-518. http://crl.ucsd.edu/~saygin/papers/MMTT.pdf
Shieber, S.M. Lessons from a Restricted Turing Test, 1993. http://www.eecs.harvard.edu/shieber/Biblio/Papers/loebner-rev-html/loebner-rev-html.html.
Whitby, B. Why The Turing Test is AI's Biggest Blind Alley, 1997. http://www.cogs.susx.ac.uk/users/blayw/tt.html

[Chinese Room]
Mooney V.J. III, Searle's Chinese Room and its Aftermath, 1997. http://csli-publications.stanford.edu/papers/CSLI-97-202.pdf
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.
Graubard, S.R. (ed.) The Artificial Intelligence Debate. False Starts, Real Foundations. Cambridge, Mass.: MIT Press. 1988.
McDermott, D. How Intelligent is Deep Blue? (1997). http://www.nyu.edu/gsas/dept/philo/courses/mindsandmachines/Papers/mcdermott.html
Minsky, M. Why People Think Computers Can't. AI Magazine, 3 No. 4 (1982). http://web.media.mit.edu/~minsky/papers/ComputersCantThink.txt.

[AI techniques]
Holland, J.H. Genetic Algorithms. (updated) http://www.econ.iastate.edu/tesfatsi/holland.GAIntro.htm.
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. ELIZA - A Computer Program For the Study of Natural Language Communication Between Man and Machine (1966). http://i5.nyu.edu/~mm64/x52.9265/january1966.html.
Winograd, T. SHRDLU. (undated). http://hci.stanford.edu/~winograd/shrdlu. (accessed 11/2008).

[AI critique]
Born, R. (ed.) Artificial Intelligence. The Case Against. London/New York: Routledge. 1988.
Dreyfus, H.L. What Computers Still Can't Do. A Critique of Artificial Reason. Cambridge, Mass.: MIT Press. 1992.
Dreyfus, H. Intelligence Without Representation (1998). http://www.class.uh.edu/cogsci/dreyfus.html.
Dreufus, H. From Socrates to Expert Systems: The Limits and Dangers of Calculative Rationality. (2004). http://socrates.berkeley.edu/~hdreyfus/html/paper_socrates.html
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/