Course Title	Minds & Machines
Course Code	AIF504
Recommended Study Year	Term 1 (Compulsory)
No. of Credits/Term	3
Mode of Tuition	Sectional approach
Class Contact Hours	3 hours of lecture and tutorial
Category	Required course
Discipline	-
Prerequisite(s)	NIL
Co-requisite(s)	NIL
Exclusion(s)	NIL
Exemption Requirement(s)	NIL

Brief Course Description:

We often think of artificial intelligence by analogy with human intelligence. Moreover, some of the most impressive machine learning systems, such as deep neural networks (DNNs), are modelled on the brain. But when, and to what extent, should we take such analogies seriously? Consider a chess-playing DNN, such as AlphaZero, for example. Does this generate representations of positions on chess boards, akin to mental representations?

And does it apply concepts such as piece mobility and space advantage, as human chess masters do, in deciding which moves to make? This course draws on the philosophy of mind and psychology to tackle such questions.

Aims:

This course aims to provide students with an overview of debates pertaining to the relationship between human intelligence and artificial intelligence. The course will focus on the relationship between human and machine intelligence, exploring the extent to which machine intelligence can be understood through analogy with human intelligence. The course will specifically focus on the use of deep neural networks (DNNs) and their relationship to human cognition.

Learning Outcomes (LOs):

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

- LO1: Accurately describe classical and contemporary theories of the mind, in the context of real and hypothetical AI systems
- LO2: Critique and evaluate classical and contemporary arguments for and against the possibility of humanlike intelligence in various real and hypothetical AIs
- LO3: Compare and contrast the organizational structures of real and hypothetical AIs, e.g. Turing machines and modular computational systems.
- LO4: Explain and evaluate analogies regarding the functioning of an AI with the functioning of the human brain.
- LO5: Evaluate arguments regarding the prospects of humanlike intelligence in near- future AI by applying philosophical theories of the mental

Indicative Content:

Week 1-3: The Turing Test, Classic approaches to AI.

Overview of the structure of the course, introduction of major themes and concepts, e.g. thought, function, intentionality, consciousness, etc., through Alan Turing's pioneering work on the possibility of artificial intelligence.

Week 4-5: The Chinese Room, Challenges for Strong AI.

Examination of John Searle's Chinese Room thought experiment, as well as other initial challenges to the Strong AI / Turing-inspired paradigm and the responses to these challenges.

Week 6-8: The "Language of Thought" and Computationalism.

Examination of computational theories of thought and mental representation, analogies and disanalogies between language and thought, introduction to basic concepts in cognitive science in connection with computationalist theories.

Week 9-11: Consciousness and emotions

Examination of the distinction between access consciousness and phenomenal consciousness, the hard problem of consciousness, the zombie argument, and the knowledge argument, comparison and contrast between consciousness and higher-order mental processes, introduction to fundamental ideas in theories of emotions and the role of emotions, if any, in AI.

Week 12: AI Models and the human brain.

Examination of AI models of brains, introduction to competing empirical hypotheses regarding the functional architecture of the brain, and existing attempts to model these possible architectures.

Week 13-14: Deep learning, AlphaZero, GPT-3, and near-future AI.

Examination of particular, existing AI systems, as well as systems that are likely to be built in the near future, with a focus on comparing these systems to the architecture of the human brain and evaluating their mentality in light of the theories discussed in previous weeks.

Teaching Method:

Lectures and discussions are aimed at explaining philosophical theories of the mind (LO1) and various mind-like artificial systems (LO3), and exploring how these views feature in arguments regarding the possibility of various kinds of artificial intelligence (LO2, LO4). Lectures and discussions are supplemented with writing assignments which assess students' facility with the relevant theories and arguments

(LO1, LO2, LO4), and a final exam that assesses students' facility with the course material generally (LO1, LO2, LO3, LO4, LO5).

Measurement of Learning Outcomes (LOs):

	Assessment Methods		
Learning Outcomes	**Class Participation and Contribution to Discussions**	**Midterm and Final Examinations**	**Final Paper**
Accurately describe classical and contemporary theories of the mind, in the context of real and hypothetical AI systems	✓	✓	✓
Critique and evaluate classical and contemporary arguments for and against the possibility of humanlike intelligence in various real and hypothetical AIs	✓	✓	✓
Compare and contrast the organizational structures of real and hypothetical AIs, e.g. Turing machines, modular computational systems, and neural nets	✓		✓
Explain and evaluate analogies regarding the functioning of an AI with the functioning of the human brain.	✓	✓	✓
Evaluate arguments regarding the prospects of humanlike intelligence in near-future AI by applying philosophical theories of the mental		✓	✓

Assessment:

Class participation and contribution to discussions: 10%

Engagement in classroom discussion and proper preparation for sessions. Students are expected to demonstrate adequate knowledge of the required weekly readings when called by the lecturer, and to critically engage with the weekly topics and discussions.

Midterm paper (1500-1800 words): 20%

Essay investigating an influential argument regarding the relationship between human intelligence and artificial intelligence (e.g., the Turing Test or Chinese Room arguments.) Should reflect the student's ability to analyze and evaluate competing theories in the philosophy of mind, and apply them to cases described in the influential arguments.

Final examination (~1.5 hours): 30%

Tests the student's knowledge of topics and readings throughout the course; e.g. short answer questions which require students to identify, describe, and contrast theories regarding the nature of the mind and mental content, as well as questions requiring students to apply these theories to existing AI systems, e.g. large language models.

Final paper (2000-2300 words): 40%

Essay evaluating the prospects of genuine mentality in existing and near-future AI systems, by applying and critically evaluating theories of the mind and mental content.

Essential Readings:

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*, 227-247.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., & Lundberg, S. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, *14*(10), e12625.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3-71.

Fodor, J. A. (1975). *The language of thought*. Thomas Y. Crowell.

Fodor, J. A. (1987). *Psychosemantics*. The MIT Press.

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly*, *32*, 127-136.

Putnam, H. (1967). Psychophysical predicates. In W. Capitan & D. Merrill (Eds.), *Art, mind, and religion* (pp. 429-440). Pittsburgh: University of Pittsburgh Press.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*, 417-457.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*, 1140-1144. https://doi.org/doi:10.1126/science.aar6404

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*, 433-460.

Weber-Guskar, E. (2021). How to feel about emotionalized artificial intelligence? When robot pets, holograms, and chatbots become affective partners. *Ethics and Information Technology*, *23*, 601-610. https://doi.org/10.1007/s10676-021-09598-8

Supplementary readings:

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., & Ji, X. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. The MIT Press.

Carruthers, P. (2006). *The architecture of the mind: Massive modularity and the flexibility of thought*. Oxford University Press.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. The MIT Press,

Hortensius, R., Hekele, F., & Cross, E. S. (2018). The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, *10*, 852-864.

Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science*, *37*, 453-488.

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 and Course Work. 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/.

Minds & Machines Final Examination (50%) Assessment Rubric

The rubric of the final exam will vary depending on the number and difficulty of the examination. The relevant rubric will be included at the beginning of the final questions.

Minds & Machines
Midterm (20%) and Final Paper (40%) Assessment Rubrics

Assessment Categories	**Excellent (A) 100–80%**	**Good (B) 79–65%**	**Pass (C) 64–50%**	**Failure (F) <50%**
Integrity (30% total)	Arguments and views relevant to the prompt are rehearsed fairly, accurately, and in sufficient detail; arguments and views are not dismissed without careful consideration.	Relevant views and arguments are rehearsed fairly, albeit with some small errors, small omissions, or unnecessary or irrelevant details.	Relevant views and arguments are accurately described, but in insufficient detail. Rehearsal of relevant views and arguments may contain substantive errors and omissions.	Relevant views or are sketched in very superficial terms; are grossly misrepresented or dismissed without argument. Paper may not address the prompt.
Rigor (30% total)	Important claims are defended with strong, developed arguments; there are clear premises that build on each other to establish the stated thesis; strong objections are anticipated and replied to.	Important claims are defended with arguments, but some premises are undersupported. Possible objections are anticipated but these objections may be developed in insufficient detail.	Some important claims are defended with arguments but others are asserted without argument; some premises are undersupported or missing; possible objections/replies are weak, obvious, or irrelevant.	The paper contains mostly assertions with little or no argument; given arguments do not support or even contradict the thesis; no possible objections are considered.
Clarity (25% total)	Clear, specific, and informative thesis; philosophical terms are appropriately introduced and defined or illustrated by example; each sentence says exactly what the author means it to say.	Thesis could be more specific; Some technical terms are used without any attempt to clarify what they mean; it is unclear what some sentences are intended to say.	There is a thesis statement but it is highly nonspecific; technical terms are often used in a way that is unclear. Some paragraphs are difficult to interpret due to problems with clarity.	No clear thesis; Much of the paper is unclear; little to no attempt is made to clarify technical terms either by definition or example.

Grammar/ Style (15% total)	Paper is very well-organized and individual paragraphs are well structured; good use of transitions between paragraphs; correct word choice. Few or no mistakes; few or no awkward constructions.	Paper has generally good overall organization but there are occasional problems with paragraph unity; some transitions unclear; some misuse of words. Occasional high-level mistakes; few or no low- level mistakes; occasional awkward constructions.	There are indications of an overall structure, but this structure is unclear and there is persistent misuse of words. Occasional lowlevel mistakes (e.g., subject-verb agreement, pronoun agreement, or spelling errors); many awkward constructions.	Overall paper structure is unclear; much of the paper cannot be assessed due to problems with grammar and style.

Unless otherwise stated, students' papers should use the latest edition of the APA (American Psychological Association) standard for citations in the midterm and final paper.

Minds & Machines
Class Participation and Contribution to Discussions (10%) Assessment Rubric

Assessment Categories	**Excellent (A) 100–80%**	**Good (B) 79–65%**	**Pass (C) 64–50%**	**Failure (F) <50%**
Class Participation and Contribution to Discussions	The student misses no more than one class without approved excuses and always contributes to the discussion by raising thoughtful questions, analyzing relevant issues, and demonstrating excellent preparation (e.g., knowledge of the assigned readings).	The student misses no more than two classes without approved excuses and sometimes contributes to the discussion in the aforementioned ways.	The student misses no more than three classes without approved excuses but only very occasionally contributes to the discussion in the aforementioned ways.	The student misses more than three classes without approved excuses and rarely or never contributes to the discussion in the aforementioned ways.