Lingnan University
Department of Philosophy

Course Title	:	Philosophy of Machine Learning
Course Code	:	PAI4003
Recommended Study Term	:	Term 1, Fourth Year
No. of Credits/Term	:	3
Mode of Tuition	:	Sectional Approach
Class Contact Hours	:	3 hours per week
Category	:	Required course
Prerequisite(s)	:	NIL
Co-requisite(s)	:	NIL
Exclusion(s)	:	NIL
Exemption Requirement(s)	:	NIL

Brief Course Description

Machine learning (ML) is indispensable to the field of Artificial Intelligence (AI), enabling the development of systems that can learn, adapt, and make intelligent decisions. Various ML-based products have been significantly integrated into our daily lives (e.g., ChatGPT, self-driving cars, recommendation systems, etc.), leading to their increasing societal impacts. To critically evaluate the social impacts and address the relevant issues, it is important to understand the fundamental concepts and techniques of ML. This course provides an introduction to these essential concepts and techniques. Additionally, it discusses various philosophical issues related to ML, including debates surrounding the limitations of ML methods and the ethical implications of applying ML technologies.

Aims

The course aims to provide students with a conceptual understanding of the current ML techniques and investigates the epistemological and ethical issues involved in the application of ML technologies.

Learning Outcomes (LOs)

Upon completion of this course, students will be able to:

1. Demonstrate a solid conceptual understanding of the key concepts and techniques of ML that are necessary for understanding the issues in the philosophical problems of ML.
2. Critically analyze the limitations of current ML methods.
3. Use the knowledge of ML to assess the main ethical problems posed by the use of ML technologies in various domains.

4. Explore the societal impacts of ML technologies.

Indicative Content

- Overview of machine learning, key concepts, and history
- Deep learning, feature engineering vs. feature learning
- Deep neural networks
- Privacy and surveillance
- Inductive reasoning and generalization
- Fairness, bias, and discrimination
- Black-box model and explainability
- The alignment problem

Teaching Method

Lecture, tutorial and experiential activities (demonstration of AI systems, chatbots, neural nets).

Measurement of Learning Outcomes

Learning Outcomes	Assessment Methods			
	Participation and in-class discussion	In-class quizzes	Midterm Paper	Final Paper
LO1: Demonstrate a solid conceptual understanding of the key concepts and techniques of ML that are necessary for understanding the issues in the philosophical problems of ML.	✓	✓	✓	✓
LO2: Critically analyze the limitations of current ML methods.	✓	✓	✓	✓
LO3: Use the knowledge of ML to assess the main ethical problems posed by	✓		✓	✓

the use of ML technologies in various domains.				
LO4: Explore the societal impacts of ML technologies.	✓		✓	✓

Assessment

There are four assessment components:

Participation and in-class discussion : 10%
In-class quizzes 10% × 2 : 20%
Midterm paper : 35%
Final paper : 35%

Students will take/write:

- Short in-class quizzes, in which they will have to answer short questions on the various concepts and theories involved in machine learning and deep learning (to assess LO1 and LO2);
- Two term papers (less than 1200 words) on two topics of the course (to assess LO1, LO2, LO3 and LO4). The students will be provided with a list of readings related to the specific topic they choose to write about.
- In addition, students are expected to participate in classroom discussions and demonstrate good preparation—for instance, they will have to demonstrate adequate knowledge of the required weekly readings when called by the lecturer (to assess LO1, LO2, LO3, and LO4).

Required/Essential Readings

Suggested:

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

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, 87(4), 568-589.

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics, 47*(5), 329-335.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16(8), e12760.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines*, 30(3), 411-437.

Karaca, K. (2021). Values and inductive risk in machine learning modelling: the case of binary classification models. *European Journal for Philosophy of Science, 11*(4), 102.

Michelfelder, D. P. (2010). Philosophy, privacy, and pervasive computing. *AI & society*, 25, 61-70.

Rieder, Gernot & Simon, Judith (2017). Big Data: A New Empiricism and its Epistemic and Socio-Political Consequences. In Wolfgang Pietsch, Jörg Wernecke and Maximilian Ott (eds.), *Berechenbarkeit der Welt?* Springer.

Russo, F., Schliesser, E., & Wagemans, J. (2023). Connecting ethics and epistemology of AI. *AI & SOCIETY*, 1-19.

Sullivan, E. (2022). Inductive risk, understanding, and opaque machine learning models. *Philosophy of Science, 89*(5), 1065-1074.

Søgaard, A. (2023). On the Opacity of Deep Neural Networks. *Canadian Journal of Philosophy*, 53(3), 224-239.

Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2016). Alignment for advanced machine learning systems. *Ethics of artificial intelligence*, 342-382.

Wheeler, Gregory (2016). Machine Epistemology and Big Data. In Lee McIntyre and Alex Rosenberg (eds.) *The Routledge Companion to Philosophy of Social Science*. Routledge.

And selections from:

Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.

Buckner, C. J. (2023). *From deep learning to rational machines: What the history of philosophy can teach us about the future of artificial intelligence*. Oxford University Press.

Coeckelbergh, M. (2020). *AI ethics*. Mit Press.

Liao, S. M. (Ed.). (2020). *Ethics of artificial intelligence*. Oxford University Press.

Kulkarni, S., & Harman, G. (2011). *An elementary introduction to statistical learning theory*. John Wiley & Sons.

Utgoff, P. E. (2012). Machine learning of inductive bias (Vol. 15). *Springer Science & Business Media*.

Other material (papers, lecture notes, etc.) will be available on Moodle.

Supplementary/Optional Readings

Suggested:

Baum, K., Mantel, S., Schmidt, E., & Speith, T. (2022). From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology, 35*(1), 12.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Desai, J., Watson, D., Wang, V., Taddeo, M., & Floridi, L. (2022). The epistemological foundations of data science: a critical review. *Synthese, 200*(6), 469.

Marmor, A. (2015). What is the right to privacy?. *Phil. & Pub. Aff.*, 43, 3.

Lyon, D. (2010). Surveillance, power and everyday life. In *Emerging digital spaces in contemporary society: Properties of technology* (pp. 107-120). London: Palgrave Macmillan UK.

Ngo, R., Chan, L., & Mindermann, S. (2022). The alignment problem from a deep learning perspective. *arXiv*:2209.00626.

Johnson, G. M. (2023). Are algorithms value-free?: Feminist theoretical virtues in machine learning. *Journal of Moral Philosophy*, 1(aop), 1-35.

And selections from:

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com

Russell, S. (2019). *Human compatible: AI and the problem of control*. Penguin Uk.

Russell, S. J., & Norvig, P. (2022). *Artificial Intelligence: A Modern Approach*, 4th, Global ed.

And some online articles, which be made available on Moodle

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 coursework, 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/.
5. Students must adhere to the University's guidelines and practices when using Generative Artificial Intelligence (AI) tools. The official documents "Guidelines for Using GAI Tools at Lingnan University" and "Best Practices for Ethical and Responsible Use of GAI Tools in Course Assessments" can be found here: https://www.ln.edu.hk/tlc/generative-artificial-intelligence/gai-guidelines-best-practices.

Philosophy of Machine Learning
Participation and in-class discussion Rubric (10%)

Assessment Categories	Excellent (A)	Good (B)	Pass (C)	Fail (F)
Participation and in-class discussion	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.

Philosophy of Machine Learning
In-class quizzes Rubric (20%)

Assessment Categories	Excellent (A)	Good (B)	Pass (C)	Fail (F)
Questions	A very good answer. The student finds the correct answer in the correct way or very nearly the correct way. Any errors are minor and does not reflect any inability to correctly use the relevant method.	An incorrect but good answer using the correct method. The correct method has been almost entirely applied in the correct way.	The correct method has been applied. The application is mostly or at least partly correct.	Either: (1) There is no answer. or (2) The wrong method was applied or the right method was applied wholly inadequately.

Philosophy of Machine Learning
Midterm paper Rubric (35%)

Assessment Categories	Excellent (A)	Good (B)	Pass (C)	Fail (F)
Integrity (30% total)	Arguments and views relevant to the prompt are rehearsed fairly, accurately, and in sufficient detail; arguments and views are not dismissed	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

	without careful consideration.			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.	Thesis is highly unspecific; 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.	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.	Overall paper structure is unclear; persistent misuse of words. Occasional low-level 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.

Philosophy of Machine Learning
Final paper Rubric (35%)

Assessment Categories	Excellent (A)	Good (B)	Pass (C)	Fail (F)
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.	Thesis is highly unspecific; 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.	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.	Overall paper structure is unclear; persistent misuse of words. Occasional low-level 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.