

Lingnan University
Department of Philosophy

Course Title	: Ethics and Artificial Intelligence
Course Code	: PAI3002
Recommended Study Term	: Term 2, Third Year
No. of Credits/Term	: 3
Mode of Tuition	: Sectional approach, with optional Service-Learning mode
Class Contact Hours	: 3 hours per week (plus 30 hours over the term for Service-Learning students)
Category	: Required course
Prerequisite(s)	: NIL
Co-requisite(s)	: NIL
Exclusion(s)	: NIL
Exemption Requirement(s)	: NIL

Brief Course Description

As Artificial Intelligence (AI) technologies continue to advance and become increasingly integrated into our daily lives, ethical questions surrounding AI are more pressing than ever. This course offers an introduction to the ethics of AI. It illustrates various ethical theories (e.g., Deontology, Consequentialism, etc.) that are useful for addressing practical issues arising from the use of AI technologies. Additionally, the course discusses a range of practical issues caused or potentially exacerbated by the widespread adoption of AI in society. Topics will cover current concerns, such as self-driving cars, deep fakes, etc., as well as future challenges related to superintelligent AI systems and existential risks.

Aims

This course aims to equip students with the necessary knowledge and methods so that they can understand, discuss, investigate, and handle the complex ethical questions raised by the introduction of AI systems into human societies. When Service-Learning is incorporated, the course aims to foster a deep interest in the harms and benefits of AI to people and communities through a variety of practical approaches.

Learning Outcomes (LOs)

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

1. Explain the key ethical theories relevant to general and specific AI ethics.
2. Identify and analyse current social and ethical problems caused by AI technologies.
3. Evaluate potential ethical dilemmas and risks associated with superintelligent or future advanced AI systems, considering short-term and long-term societal impacts.
4. **Non-Service-Learning mode:** Critically analyse and propose solutions to AI-related ethical problems by synthesising knowledge of AI systems, technologies and ethical theories.

Service-Learning mode: Evaluate the real-world impacts of AI on people and communities by assessing harms and benefits, then designing actionable recommendations for community partners, organisations, or policymakers to address these issues effectively.

Indicative Content

- Introduction to ethical theories: consequentialism, deontology, virtue ethics, etc.
- Algorithmic bias.
- Black-box algorithms.
- Self-driving cars.
- Autonomous weapon systems.
- AI companions.
- Griefbots and
- Sexbots.
- Deep fake and misinformation.
- Labour displacement.
- Superintelligent AI and existential risk.

Teaching Method

All students, both with and without Service-Learning component, will attend an initial 10 weeks of classes. This period will address key topics in AI ethics, including foundational concepts such as algorithmic bias and opaque decision-making, as well as applied AI ethics issues such as the ethics of self-driving cars, AI companions, and labour displacement. Students will take a test after each section of the course, which will assess students' learning of concepts through questions that involve short answers.

Subsequently, students will be divided into two groups: the Service-Learning group and the non-Service-Learning group, based on the students' preferences. The Service-Learning group will work on their Service-Learning project (further details provided below). Meanwhile, the non-Service-Learning Group will attend further workshops on essay-writing, preparing them for the research essay on their chosen topic. In the final 2 weeks, each student will give a 10-minute presentation based on their initial draft of their research essay, followed by a 5-minute Q&A session. This will provide students with the opportunity to receive feedback from their peers, which they may draw upon before submitting their final essay of up to 1,500 words.

Service-Learning project:

Students will be divided into groups for Service-Learning Project. Students will be required to spend 30 hours applying the knowledge learned from lectures about the potential harms and benefits of AI systems and transfer the knowledge to community partners. The course instructor and the staff from the Office of Service-Learning (OSL) will work together and decide the activity for students each semester. In addition, students will receive training from the OSL

staff to work with the community partner. Further information about the Service-Learning projects will be provided by OSL in the first lecture of the term.

Some examples of potential projects:

- **AI auditing projects:** Students partner with an organisation to examine a tool or system they use such as a hiring algorithm, a chatbot, or a content moderation system, and produce an ethics audit report covering fairness, transparency, and accountability. The deliverable goes to the organisation as a usable document.
- **Community AI literacy workshops:** Students design and facilitate workshops for community members (seniors, jobseekers, students, etc.) on topics like how recommendation algorithms work, how to recognize AI-generated content, or what rights they have when AI is used in decisions about them. This requires students to translate complex theory into accessible language.

In addition to the activity, students are required to prepare a short presentation before the Service-Learning activity. Each group will present their ideas for the activity, including methods that they will be using transfer knowledge to community partners effectively. Students from other groups are encouraged to provide feedback and comments on each group's work, which they may choose to take on when executing the activity. Students will be graded on the quality of their ideas and the aptitude with which they respond to other students' questions. After the Service-Learning activity, each student is required to submit a written report of up to 1,500 words, describing the work that was done by the group and what they have learnt from the Service-Learning project.

Measurement of Learning Outcomes

Without Service-Learning

Learning outcomes	Assessment methods			
	<i>Participation and in-class discussion</i>	<i>In-class tests (x4)</i>	<i>In-class presentation</i>	<i>Research essay</i>
LO1: Explain the key ethical theories relevant to general and specific AI ethics.	X	X	X	X
LO2: Identify and analyse current social and ethical problems caused by AI technologies.	X	X	X	X
LO3: Evaluate potential ethical dilemmas and risks associated with superintelligent or future advanced AI systems, considering short-term and long-term societal impacts.	X	X	X	X

LO4: Critically analyse and propose solutions to AI-related ethical problems by synthesising knowledge of AI systems, technologies and ethical theories.	X	X	X	X
--	---	---	---	---

With Service-Learning

Learning outcomes	Assessment methods		
	<i>Participation and in-class discussion</i>	<i>In-class tests (x4)</i>	<i>Service-Learning project</i>
LO1: Explain the key ethical theories relevant to general and specific AI ethics.	X	X	
LO2: Identify and analyse current social and ethical problems caused by AI technologies.	X	X	X
LO3: Evaluate potential ethical dilemmas and risks associated with superintelligent or future advanced AI systems, considering short-term and long-term societal impacts.	X	X	X
LO4: Evaluate the real-world impacts of AI on people and communities by assessing harms and benefits, then designing actionable recommendations for community partners, organisations, or policymakers to address these issues effectively.			X

Assessment

Without Service-Learning

Participation and in-class discussion	: 10%
In-class tests (10% x 4)	: 40%
In-class presentation	: 15%
Research essay	: 35%

With Service-Learning

Participation and in-class discussion	: 10%
In-class tests (10% x 4)	: 40%
Service-Learning Project	
Presentation	: 15%
Service-Learning report	: 35%

Required/Essential Readings

Suggested:

- Johnson, G. M. (2020). Algorithmic bias: On the implicit biases of social technology. *Synthese*, 198(10): 9941–9961. <https://doi.org/10.1007/s11229-020-02696-y>
- Vredenburg, K. (2022). The Right to Explanation*. *Journal of Political Philosophy*, 30: 209-229. <https://doi.org/10.1111/jopp.12262>
- Francis, T., Karhu, T (2025). Getting machines to do your dirty work. *Philosophical Studies* 182, 121–135. <https://doi.org/10.1007/s11098-023-02027-0>
- Nyholm S. (2018). The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, 13:e12507. <https://doi.org/10.1111/phc3.12507>
- Kauppinen, A. (2020). Who should bear the risk when self-driving vehicles crash? *Journal of Applied Philosophy*, 38: 630-645. <https://doi.org/10.1111/japp.12490>
- Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, 24: 62-77. <https://doi.org/10.1111/j.1468-5930.2007.00346.x>
- Danaher, J. (2019). The philosophical case for robot friendship. *Journal of Posthuman Studies*, 3(1):5–24. <https://doi.org/10.5325/jpoststud.3.1.0005>
- Sung, L. U., & Ferdman, A. (2025). Empathetic Large Language Models, the social capacities and human flourishing. *Inquiry*, 1–23. <https://doi.org/10.1080/0020174X.2025.2518448>
- Lindemann, N.F. The Ethics of ‘Deathbots’. *Sci Eng Ethics* 28, 60 (2022). <https://doi.org/10.1007/s11948-022-00417-x>
- Esken, R. (2017). Is sex with robots rape? *Journal of Practical Ethics*, 5(2), 62. <http://www.jpe.ox.ac.uk/wp-content/uploads/2017/12/JPE0042-Eskens-2.pdf>
- Rini, R. (2020). Deepfakes and the epistemic backstop. *Philosophers’ Imprint*, 20 (24):1-16. <http://hdl.handle.net/2027/spo.3521354.0020.024>.
- Rini, R. and Cohen, L. (2022). Deepfakes, deep harms. *Journal of Ethics and Social Philosophy* 22 (2): 143 – 161. <https://doi.org/10.26556/jesp.v22i2.1628>.
- Gheaus, A. & Herzog, L. (2016). The goods of work (other than money!). *Journal of Social Philosophy*, 47: 70-89. <https://doi.org/10.1111/josp.12140>.
- Vold, Karina & Harris, Daniel (2021). How does artificial intelligence pose an existential risk? *The Oxford Handbook of Digital Ethics*. <https://doi.org/10.1093/oxfordhb/9780198857815.013.36>

Supplementary/Optional Readings

Suggested:

- Etzioni, A., & Etzioni, O. (2016). AI assisted ethics. *Ethics and Information Technology*. <https://doi.org/10.1007/s10676-016-9400-6>.
- Floridi, L. (2016). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, 374(2083), 1–13.
- Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*. <https://doi.org/10.1007/s11023-020-09517-8>.
- Hooker, J., & Hooker, J. (2018). Ethics of Artificial Intelligence. In *Taking Ethics Seriously*. <https://doi.org/10.4324/9781315097961-14>.
- McDermott, D. (2008). Why ethics is a high hurdle for AI. *North American Conference on Computers and Philosophy (NA-CAP)*.
- Mittelstadt, B. (2019). AI Ethics – Too Principled to Fail? *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3391293>.
- Sparrow, R. (2007). Killer robots. *Journal of applied philosophy*, 24(1), 62-77.

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/cital/generative-artificial-intelligence/gai-guidelines-best-practices>.

Assessment Rubrics

For both Service-Learning and Non-Service-Learning students:

Participation and in-class discussion (10%)

Assessment Categories	Excellent (A, A-)	Good (B+, B, B-)	Fair (C+, C, C-)	Pass (D+, D)	Failure (F)
Participation and in-class discussions	Misses no more than one class without approved excuses. Frequently contributes to the discussion by raising thoughtful questions/comments.	Misses no more than two classes without approved excuses. Sometimes contributes to the discussion in the aforementioned ways.	Misses no more than three classes without approved excuses. Only occasionally contributes to the discussion in the aforementioned ways.	Misses no more than five classes without approved excuses. Rarely or never contributes to the discussion in the aforementioned ways.	The student misses more than five of the classes without approved excuses. Rarely or never contributes to the discussion in the aforementioned ways.

In-class tests (4 x 10% = 40%)

Assessment Categories	Excellent (A, A-)	Good (B+, B, B-)	Fair (C+, C, C-)	Pass (D+, D)	Failure (F)
Accuracy, critical engagement & clarity	Answers show an excellent understanding of the content, using precise and clear writing. Student demonstrates strong ability to critically engage with the material.	Answers show a good understanding of the content, but with minor mistakes or lack of detail. Student demonstrates a reasonable ability to critically engage with the material.	Answers show a partial understanding of the content, with some mistakes or significant lack of detail. There is some critical engagement but lacks depth.	Answers show a poor understanding of the content, with major mistakes and little to no critical engagement.	The student either fails to answer the question entirely or demonstrates little to no understanding of the content. Writing is unclear, and there is no critical engagement with the material.

For Non-Service-Learning students ONLY:

In-class presentation (15%)

Assessment Categories	Excellent (A, A-)	Good (B+, B, B-)	Fair (C+, C, C-)	Pass (D+, D)	Failure (F)
Subject knowledge (40%)	Demonstrates full knowledge with clear detailed explanations. Able to answer questions with aptitude.	Demonstrates good knowledge with explanations. Answers to questions are adequate but lacking in detail and depth.	Demonstrates fair knowledge. Can only answer rudimentary questions about the subject.	Limited information. Can hardly answer questions about the subject.	No grasp of information and cannot answer questions about the subject.

Critical engagement (30%)	Demonstrates a strong ability to critically engage with the literature; shows a significant amount of independent thought/originality, which is reflected in both the presentation and answers to peers' questions.	Demonstrates a reasonable ability to critically engage with the literature; shows some independent thought/originality, which is reflected in both the presentation and answers to peers' questions.	Demonstrates some ability to critically engage with the literature; there is a lack of independent thought, with the student mostly regurgitating class material.	There is very little evidence of critical engagement with the literature, and no evidence of independent thought.	There is no evidence of any critical engagement with the literature. For example, there is no reference to the views discussed in class, and there are no objections/counter examples which are considered.
Presentation/communication skills (30%)	Well-prepared visual aids; clear, loud voices, engage with the audience, smooth delivery.	Satisfactory visual aids; voice generally clear, some engagement with the audience, generally smooth delivery.	Fair visual aids; voice too low, minimal engagement of audience, fair delivery.	Insufficient visual aids; voice unclear, no engagement with audience, poor delivery.	Little to no visual aids, voice cannot be heard, no engagement with audience, poor delivery.

Research essay (35%)

Assessment Categories	Excellent (A, A-)	Good (B+, B, B-)	Fair (C+, C, C-)	Pass (D+, D)	Failure (F)
Integrity (35%)	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 superficial terms; are misrepresented or dismissed without argument.	Paper fails to address the prompt in any meaningful way, demonstrating little to no understanding of the relevant views and arguments.
Rigour (35%)	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 under-supported. 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 under-supported or missing; possible objections/replies are weak, obvious, or irrelevant.	There is some attempt at defending claims made in the paper but contains mostly assertions with little or no argument.	The paper contains arguments which do not support or even contradict the thesis; no possible objections are considered.
Clarity (20%)	Clear, specific, and informative thesis;	Thesis could be more specific; Most	Thesis is unspecific; Philosophical/tech	Thesis is highly unspecific; philosophical/tech	No clear thesis; Much of the paper is unclear; little to

	philosophical/technical terms are appropriately introduced and defined or illustrated by example; each sentence says exactly what the author means it to say.	philosophical/technical terms are appropriately introduced and defined; it is unclear what some sentences are intended to say.	nical terms are largely used without appropriately introducing them; it is unclear what a significant number of sentences are intended to say.	nical terms are used in a way that is unclear. Some paragraphs are difficult to interpret due to problems with clarity.	no attempt is made to clarify philosophical/technical terms either by definition or example.
Grammar/Style (10%)	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; substantial misuse of words. Occasional low-level mistakes (e.g., subject verb agreement, pronoun agreement, or spelling errors); awkward constructions.	Overall paper structure is unclear; persistent misuse of words. Frequent 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.

For Service-Learning students ONLY:

Service-Learning project presentation (15%)

Assessment Categories	Excellent (A, A-)	Good (B+, B, B-)	Fair (C+, C, C-)	Pass (D+, D)	Failure (F)
Content and community context (40%)	Demonstrates evidence-based portrait of community and integration of partner voices. Goals, activity, and intended outcomes are presented with precision.	There is a clear overview of the community needs with supporting evidence. Goals and intended outcomes are clearly communicated.	There is a basic description of community context, but limited evidence or partner perspective. Outcomes are described but goals are not clearly linked to activities.	Community context is vague or inaccurate. Scope and outcomes are not very clear, with little evidence to relevance to activities.	Little to no evidence of community context. Scope and outcomes not stated, unclear, or not relevant to the activities.
Academic integration (30%)	Explicit connections to course concepts, and clear evidence of how theory actively illuminates the project. This is demonstrated	There are clear connections to relevant theory, which is reflected in both the presentation and answers to peers' questions.	Some theory is referenced but connections are superficial or underdeveloped. Claims are not always grounded in evidence, and students fail to	There is little integration of academic content with service activities. Sources may be provided but are irrelevant with	No evidence of integration with academic content with service activities. Sources entirely absent.

	both in the presentation and answers to peers' questions.		adequately explain theory-practice connection in Q&A.	unsubstantiated claims made.	
Presentation/communication skills (30%)	Well-prepared visual aids; clear, loud voices, engage with the audience, smooth delivery.	Satisfactory visual aids; voice generally clear, some engagement with the audience, generally smooth delivery.	Fair visual aids; voice too low, minimal engagement of audience, fair delivery.	Insufficient visual aids; voice unclear, no engagement with audience, poor delivery.	Little to no visual aids, voice cannot be heard, no engagement with audience, poor delivery.

Service-Learning project report (35%)

Assessment Categories	Excellent (A, A-)	Good (B+, B, B-)	Fair (C+, C, C-)	Pass (D+, D)	Failure (F)
Content (30%)	All the information included is highly relevant to the issue being addressed.	Most of the information included is relevant to the issue being addressed.	A sufficient proportion of the information included is relevant to the issue being addressed.	Much of the information included is irrelevant, or is of questionable relevance to the issue being addressed.	Report fails entirely to provide relevant information to the issue being addressed.
Critical reflection (40%)	Demonstrates transformative learning through depth of personal reflection; critically interrogates assumptions, systemic inequalities, and the students' own role'; engages with any ethical tensions honestly.	Demonstrates thoughtful reflection on learning and values, going beyond description to analysis. Student identifies key ethical issues and engages meaningfully with them.	There is some descriptive reflection present, but a limited analysis of how the experience challenged thinking. There is an acknowledgement of ethical issues but fails to engage meaningfully with them.	Reflection is superficial, with the report mostly describing what activity was done without meaningful insight. Ethics not addressed or addressed only at a surface level.	No evidence of any critical reflection or identification of ethical issues.
Clarity (20%)	Report is clear, specific, and informative, with no gaps, ambiguities, or inaccuracies. Any philosophical/technical terms are appropriately introduced, and each sentence says exactly what the author means it to say.	Ideas, points, and explanations in the report are clear for the most part. Most philosophical/technical terms are appropriately introduced and defined; it is unclear what some sentences are intended to say.	Ideas, points, and explanations in the report tend to lack sufficient clarity. Philosophical/technical terms are largely used without appropriately introducing them; it is unclear what a significant number of	Report is highly unspecific; philosophical/technical terms are used in a way that is unclear. Some paragraphs are difficult to interpret due to problems with clarity. There are significant gaps, ambiguities, and inaccuracies.	Report is entirely unclear, and there is little to no attempt is made to clarify philosophical/technical terms either by definition or example.

			sentences are intended to say.		
Grammar/Style (10%)	Report 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 but generally easy to follow; some transitions unclear. The quality of writing is good with few, though noticeable, errors in word usage.	Report structure generally unclear; substantial misuse of words. Occasional low-level mistakes (e.g., subject verb agreement, pronoun agreement, or spelling errors); awkward constructions.	Overall report structure is unclear and difficult to follow; persistent misuse of words. Frequent low-level mistakes (e.g., subject verb agreement, pronoun agreement, or spelling errors); many awkward constructions.	Report is structured in a way that makes it very difficult to follow; much of the paper cannot be assessed due to problems with grammar and style.