THE UNIVERSITY OF HONG KONG

The Philosophy of Artificial Intelligence

PHIL2225

Academic Year 2024-2025, Term I

Instructor : David Villena (dvillena@hku.hk)

Lectures : Thursdays, 10:30-12:20

Classroom : Centennial Campus CPD-2.37

Prerequisites : None

Credits : 6.0

Course abstract

This course provides an introduction to the philosophy of artificial intelligence. Central questions include: In what sense is artificial intelligence a form of intelligence? Can computers achieve a form of intelligence that far surpasses human intelligence, i.e., can we reach the so-called Singularity? Does the Singularity pose an existential threat to humanity? Can computers be conscious? Can we communicate and share a language with AI? Can and should decisions made by AI be explainable to humans? What is algorithmic bias and what can be done about it? What are our moral obligations towards AI?

Learning outcomes

On successful completion of this course, students should be able to:

- Demonstrate an understanding of the argument structure and norms of argumentation specific to debates in the philosophy of artificial intelligence, and critically evaluate key arguments within this domain.
- Construct persuasive arguments that address foundational questions in the philosophy of artificial intelligence, such as the nature of intelligence, consciousness, language, and the ethical implications of AI.
- Present and critically evaluate ideas and arguments related to the philosophy of artificial intelligence, identifying and exploring alternative perspectives.
- Articulate their own views on the philosophical implications of AI, developing insights into fundamental questions about consciousness, language, understanding, agency, and value.
- Engage in constructive, open-minded discussions with others who hold different views on AI, presenting their own positions clearly and logically.
- Appreciate the range and variety of philosophical projects related to AI, understanding how these perspectives have evolved throughout history and across different societies.
- Use philosophical skills to interpret and understand the views of others concerning artificial intelligence, expressing their own views clearly, concisely, and effectively both orally and in writing.
- Appreciate the role of open, rational discussion in the shared pursuit of knowledge about artificial intelligence and its implications for society.

Assessment methods

- Midterm test (20%)
- Final test (20%)
- Group presentation 1 (15%)
    - Presentation and critical evaluation of an influential thesis and argument related to the philosophy of artificial intelligence.
- Group presentation 2 (20%)

    - Presentation and defense of group's personal view and argument related to the philosophy of artificial intelligence.
- Peer review and feedback on group presentations (15%)
- Short paper (1200 words not including references) (10%)

To pass this course, students are required to fulfill all components of the assessment methods.

Grading scale

≥ 90%	A+	≤ 100%	Excellent
≥ 85%	A	< 90%	
≥ 80%	A-	< 85%	
≥ 75%	B+	< 80%	Good
≥ 70%	B	< 75%	
≥ 65%	B-	< 70%	
≥ 60%	C+	< 65%	Satisfactory
≥ 55%	C	< 60%	
≥ 50%	C-	< 55%	
≥ 45%	D+	< 50%	Pass
≥ 40%	D	< 45%	
0%	F	< 40%	Fail

Academic honesty

Plagiarism is a serious academic offense. Coursework submitted for assessment purposes must be students' own work and properly acknowledge all sources. Students who plagiarize might be referred to the University's disciplinary committee and might be expelled from the University as a result. You can find more information at http://www.hku.hk/plagiarism. If you have any doubts as to whether the work you are handing in constitutes plagiarism, please don't hesitate to consult with the instructor beforehand.

Disabilities

Inform the instructor if you have a visible or non-visible disability (e.g., a mental health condition) that may have some impact on your work in this class, so reasonable accommodations and assistance can be provided to you.

Screen recommendations

It would be greatly appreciated if you could minimize the use of your phone and other screen devices like laptops, tablets, smartwatches, and e-readers during class. If you have an emergency or expect an important call, please inform the instructor.

Additionally, you do not need to take pictures of the slides, as they will be available to download on Moodle. While the use of laptops is not completely discouraged, we kindly ask that you use them primarily for note-taking purposes. This is because multitasking, like surfing the internet or emailing during class, can potentially distract you and those around you in the classroom. We understand that everyone has different learning styles, so we just ask that you consider these recommendations for the benefit of both your learning experience and that of your classmates.

Course contents and materials

Unit I. Mind, language, and metaphysics

1. Introduction, multiple realizability, and consciousness

Suggested readings:

- Chalmers, D. J. (2023). Could a large language model be conscious? *arXiv preprint*. https://arxiv.org/abs/2303.07103
- Lemoine, B. (2022a, June 11). Is LAMDA sentient? An interview. https://cajundiscordian.medium.com/is-lamda-sentient-an-interview-ea64d916d917
- Lemoine, B. (2022b, June 11), What is LaMDA and what does it want? https://cajundiscordian.medium.com/what-is-lamda-and-what-does-it-want-688632134489
- Lemoine, B. (2022c, August 15). What is sentience and why does it matter? *Medium*.https://cajundiscordian.medium.com/what-is-sentience-and-why-does-it-mater-2c28f4882cb9
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450.

Suggested viewing:

- Jonze, S. (Director). (2013). *Her* [Film]. Annapuma pictures.

Supplementary readings:

- Block, N. (2009). Comparing the major theories of consciousness. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences*, 4th edition (pp. 1111-1122). The MIT Press.
- Butlin, P., et al. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint*. https://arxiv.org/abs/2308.08708
- Chalmers, D. J. (1995). Facing up the problem of consciousness. *Journal of Consciousness Studies*, *2*, 200-219.

- Dehane, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*, 486-492.
- Lenharo, M. (2023, June 24). Decades-long bet on consciousness ends—and it's philosopher 1, neuroscientist 0. *Nature*. https://www.nature.com/articles/d41586-023-02120-8
- Levy, S. (2022, June 17). Blake Lemoine says Google's LaMDA AI faces 'bigotry'. *Wired*.https://www.wired.com/story/blake-lemoine-google-lamda-ai-bigotry/
- Schneider, S. (2014, March 2). The philosophy of 'Her.' *The New York Times*. https://archive.nytimes.com/opinionator.blogs.nytimes.com/2014/03/02/the-philosophy-of-her/
- Schneider, S. (2019). *Artificial you: AI and the future of your mind*. Princeton University Press. (Chapter 2, The problem of AI consciousness; Chapter 3, Consciousness engineering; Chapter 4, How to catch an AI zombie: Testing for consciousness in machines)
- Shanahan, M. (2016, October 19). Conscious exotica. *Aeon*. https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there
- Sunday Grevè, S., & Yu, X. Can machines be conscious? (2023). *Philosophy Now*, 155. https://philosophynow.org/issues/155/Can_Machines_Be_Conscious
- Tiku, N. (2022, June 11). The Google engineer who thinks the company's AI has come to life. *The Washington Post*. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

2. Artificial Intelligence & Artificial General Intelligence: Turing Test and the Chinese Room Argument

Suggested readings:

- Boden, M. A. (2016). *AI: Its nature and future*. Oxford University Press. (Chapter 1, What is artificial intelligence?)
- Bubeck, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint*. https://arxiv.org/abs/2303.12712

- Cappelen, H. & Dever, J. (2021). *Making AI Intelligible: Philosophical foundations*. Oxford University Press. (Chapter 4, Our theory: De-anthropocentrized externalism)
- Russell S., & Norvig, P. (2020). *Artificial intelligence: A modern approach*, 4th edition. Pearson. (Chapter 1, Introduction)
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*, 417-424.
- Turing, A. M. (1950). Computing machinery and intelligence, *Mind*, *59*, 433-460.

Suggested viewing:

- Bubeck, S. (2023, April 7). Sparks of AGI: Early experiments with GPT-4. Talk. (48m 31s)
- Kohls, G. (Director). (2017) AlphaGo [Documentary]. Moxie Pictures, Google DeepMind.

Supplementary readings:

- Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, *14*: e12625. https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/phc3.12625
- Cole, D. (2023). The Chinese Room Argument. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition). https://plato.stanford.edu/archives/sum2023/entries/chinese-room/
- DeBellis, M. (2023). Arguing with the Chinese Room. *Philosophy Now*, 155. https://philosophynow.org/issues/155/Arguing_with_the_Chinese_Room
- Flasiński, M. (2016). *Introduction to artificial intelligence*. Springer. (Chapter 1, History of artificial intelligence)
- Hernández-Orallo, J., & Dowe, D. L. (2010). Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, *174*, 1508-1539.
- Lanier, J. (2023, April 20). There is no A. I. *The New Yorker*. https://www.newyorker.com/science/annals-of-artificial-intelligence/there-is-no-ai

- Millière, R., & Buckner, C. (2024). A philosophical introduction to language models--Part I: Continuity with classic debates. *arXiv preprint* https://arxiv.org/abs/2401.03910
- Mitchell, M. (2023) How do we know how smart AI systems are? *Science*, *381*, adj5957. https://www.science.org/doi/10.1126/science.adj5957
- Sunday Grève, S. (2022, April 22). AI's first philosopher. *Aeon*. https://aeon.co/essays/why-we-should-remember-alan-turing-as-a-philosopher

Unit II. Epistemology, agency, and value

3. AI, beliefs, testimony, and emotions

Suggested readings:

- Dennett, D. C. (2023, May 16). The problem with counterfeit people. *The Atlantic*.https://www.theatlantic.com/technology/archive/2023/05/problem-counterfeit-people/674075/
- Park, P. S., et al. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, *5*(5). https://doi.org/10.1016/j.patter.2024.100988
- Pepp, J., Michaelson, E., & Sterken, R. K. (2019). What's new about fake news? *Journal of Ethics and Social Philosophy*, *16*, 67-94.
- Roose, K. (2023, February 16a). A conversation with Bing's chatbot left me deeply unsettled. *The New York Times*. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html
- Vaidya, A. J. (2024). Can machines have emotions? *AI & Society*. https://doi.org/10.1007/s00146-024-02022-x
- Villena, D. (2022). Deepfakes, deception, and distrust: Epistemic and social concerns. *Daily Philosophy.* https://daily-philosophy.com/david-villena-deepfakes/

Suggested viewing:

- Garland, A. (Director). (2014). *Ex machina* [Film]. Film4 & DNA Films.

Supplementary readings:

- El Atillah, I. (2023, March 31). Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change. *Euronews*. https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-
- Knight, W. (2023, May 24). These ChatGPT rivals are designed to play with your emotions. *Wired*.https://www.wired.com/story/fast-forward-chatgpt-rivals-emotions/
- Roose, K. (2023, February 16b). Bing's A.I. chat: 'I Want to Be Alive. 😈'. *The New York Times*. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html
- Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2023, May 29). Creating a large language model of a philosopher. Preprint. http://www.faculty.ucr.edu/~eschwitz/SchwitzPapers/GPT-3-Dennett-230509.pdf
- Sterri, A. B., & Earp, B. D. (2021). The ethics of sex robots. In C. Véliz (Ed.), *The Oxford handbook of digital ethics* (Online edition). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198857815.013.13

4. AI ethics and algorithmic bias

Suggested readings:

- Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. Ramsey (Eds.), *The Cambridge handbook of Artificial Intelligence* (pp. 316-334). Cambridge University Press.
- Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, *16*, e12760. https://doi.org/10.1111/phc3.12760
- Kasirzadeh, A. (2021). Reasons, values, stakeholders: A philosophical framework for explainable artificial intelligence. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3442188.3445866

- Liao, S. M. (2020). The moral status and rights of artificial intelligence. In S. M. Liao (Ed.), *Ethics of artificial intelligence* (pp. 480-503). Oxford University Press.
- Vredenburgh, K. (2024). Fairness. In J. B. Bullock et al. (Eds.), *The Oxford Handbook of AI Governance* (pp. 129-148). Oxford University Press.

Suggested viewing:

- Snodgrass, M. M. (Writer), & Scheerer, R. (Director). (1989, February 13). The measure of a man (Season 2, Episode 9) [TV series episode]. In *Star Trek: The Next Generation*. Paramount Domestic Television.

Supplementary readings:

- Angwin, J., et al. (2016, May 23). Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- Awad, E., et al. (2018). The Moral Machine experiment. *Nature, 563*, 59-64.
- Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, *81*, 149-159.
- Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*, 10025.
- Krügel, S., Ostermaier, A. & Uhl, M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. *Nature Scientific Reports,* 13, 4569. https://doi.org/10.1038/s41598-023-31341-0
- Schwitzgebel, E. and Garza, M. (2015), Defense of the rights of artificial intelligences. *Midwest Studies In Philosophy*, *39*, 98-119
- Schwitzgebel, E., & Garza, M. (2020). Designing AI with rights, consciousness, self-respect, and freedom. In S. M. Liao (Ed.), *Ethics of artificial intelligence* (pp. 459-479). Oxford University Press.
- Schwitzgebel, E. (2023, March 16). Don't create AI systems of disputable moral status. *Daily Nous*. https://dailynous.com/2023/03/16/philosophers-on-next-generation-large-language-models/#schwitzgebel

5. The ethics of autonomous weapons

Suggested readings:

- Asaro, P. (2020). Autonomous weapons and the ethics of artificial intelligence. In S. M. Liao (Ed.), *Ethics of artificial intelligence* (pp. 212-236). Oxford University Press.
- Leveringhaus, A. (2016). *Ethics and autonomous weapons*. Palgrave. (Chapter 1, Ethics and the autonomous weapons debate; Chapter 2, Autonomous weaponry: Conceptual issues)
- Sparrow, R. (2007). Killer robots. Journal of Applied Philosophy, 24, 63-77.
- Sparrow, R. (2016). Robots and respect: Assessing the case against autonomous weapon systems. *Ethics & International Affairs*, *30*, 93-116.
- Steinhoff, U. (2013). Killing them safely: Extreme asymmetry and its discontents. In B. J. Strawser (Ed.), *Killing by remote control: The ethics of an unmanned military* (pp. 179-208). Oxford University Press.
- Taddeo, M., & Blanchard, A. (2022). A comparative analysis of the definitions of autonomous weapons systems. *Science and Engineering Ethics*, 28, 37.

Supplementary readings:

- Anderson, K., & Waxman, M. C. (2013). *Law and ethics for autonomous weapon systems: Why a ban won't work and how the laws of war can*. Stanford University, The Hoover Institution Jean Perkins Task Force on National Security & Law Essay Series. https://scholarship.law.columbia.edu/faculty_scholarship/1803
- Galliott, J. (2021). Toward a positive statement of ethical principles for military AI. In J. Galliott, D. MacIntosh, & J. D. Ohlin (Eds.), *Lethal autonomous weapons: Re-examining the law and ethics of robotic warfare* (pp. 121-136). Oxford University Press.
- Leveringhaus, A. (2018). What's so bad about killer robots? *Journal of Applied Philosophy*, *35*, 341-358.
- Lucas, G. (2023). *Law, ethics and emerging military technologies: Confronting disruptive innovation*. Routledge. (Chapter 3, Ethics and automated warfare)
- Purves, D., Jenkins, R., & Strawser, B. J. (2015). Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice*, *18*, 851-872.

- Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. W. W. Norton & Company. (Chapter 3, Machines that kill: What is an autonomous weapon?; Chapter 17: Soullles killers: The morality of autonomous weapons)
- Simpson, T. W., & Müller, V. C. (2016). Just war and robots' killings. *The Philosophical Quarterly*, *66*, 302-322.
- Singer, P. W. (2009). *Wired for war: The robotics revolution and conflict in the twenty-first century*. Penguin. (Chapter 6, Always in the loop? The arming and autonomy of robots)
- Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, *24*, 62-77.
- Wood, N. G. (2023). Autonomous weapon systems: A clarification. *Journal of Military Ethics,* 22(1), 18-32.

7. Creativity and AI

Suggested readings:

- Boden, M. A. (1998). Creativity and artificial intelligence. *Artificial Intelligence*, *103*(1-2), 347-356.
- Boden, M. A. (2007). Creativity in a nutshell. *Think*, *5*(15), 83-96.
- Livingston, P. (2018). Explicating "creativity". In B. Gaut & M. Kieran (Eds.), *Routledge handbook on creativity and philosophy* (pp. 108-123). Routledge.

Suggested viewings:

- Rubin, J. (Director). (2023). *The frost* [AI-generated film]. Waymark Creative Lab & Latent Cinema.
- *The Economist*. (2021, April 7). How AI is transforming the creative industries. (8m 26s)
- The Museum of Modern Art (MoMA). (2023, March 16). AI Art: How artists are using and confronting machine learning. How to see like a machine. (14m 56s)
- Thomas, Z. (2023, April 18). How AI could replace artists in creative industries. *The Wall Street Journal*. (16m 38s)
- *Wired*. (2020, January 17). How this guy uses A.I. to create art. (10m 32s)

Supplementary readings:

- Eisikovits, N., & Stubbs, A. (2023, January 12). ChatGPT, DALL-E 2 and the collapse of the creative process. *The Conversation*. https://theconversation.com/chatgpt-dall-e-2-and-the-collapse-of-the-creative-process-196461
- Gaut, B. (2010). The philosophy of creativity. *Philosophy Compass*, *5*, 1034-1046.
- Grierson, J. (2023, April 17). Photographer admits prize-winning image was AI-generated. *The Guardian*. https://www.theguardian.com/technology/2023/apr/17/photographer-admits-prize-winning-image-was-ai-generated
- Harwell, D. (2022, September 2). He used AI to win a fine-arts competition. Was it cheating? *The Washington Post*. https://www.washingtonpost.com/technology/2022/09/02/midjourney-artificial-intelligence-state-fair-colorado/
- Kelly, K. (2022, November 17). Picture limitless creativity at your fingertips. *Wired*.https://www.wired.com/story/picture-limitless-creativity-ai-image-generators/
- Nguyen, C. T. (2020, July 30) Who trains the machine artist? *Daily Nous*. https://dailynous.com/2020/07/30/philosophers-gpt-3/#nguyen
- Picciuto, E., & Carruthers, P. (2014). The origins of creativity. In E. S. Paul & S. B. Kaufman (Eds.),*The philosophy of creativity: New essays* (pp. 199-223). Oxford University Press.
- Roose, K. (2022, September 2). An A.I.-generated picture won an art prize. Artists aren't happy. *The New York Times*. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html
- Shepherd, T. (2023, July 11). Woman's iPhone photo of son rejected from Sydney competition after judges ruled it could be AI. *The Guardian*. https://www.theguardian.com/australia-news/2023/jul/11/mothers-iphone-photo-of-son-rejected-from-sydney-competition-after-judges-ruled-it-could-be-ai
- Shevlin, H. (2021). Rethinking creative intelligence: Comparative psychology and the concept of creativity. *European Journal for Philosophy of Science*, *11*, Article number: 16.

Unit III. Synthesis

8. Artificial Superintelligence, alignment, and the Singularity

Suggested readings:

- Boden, M. A. (2016). *AI: Its nature and future*. Oxford University Press. (Chapter 7, The singularity)
- Bostrom, N. (2003). Are we living in a computer simulation? *Philosophical Quarterly*, *53*, 243-255.
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press. (Chapter 9, The control problem)
- Chalmers, D. J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies, 17*(9-10), 7–65.
- Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds & Machines*, *30*, 411–437.

Suggested viewings:

- Bengio, Y., Tegmark, M., Mitchell M., & LeCun, Y. (2023, June 24). Munk Debate on Artificial Intelligence. Be it resolved, AI research and development poses an existential threat. (1h 47m 19s)
- Bostrom, N. (2015). What happens when our computers get smarter than we are? TED Talk. https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are(16m 22s)
- University of Toronto. (2023, June 23). The Godfather in Conversation: Why Geoffrey Hinton is worried about the future of AI. (46m 20s)
- Yudkowsky, E. (2023, July 12). Will superintelligent AI end the world? TED Talk. (10m 32s)

Suggested listening:

- Harris, S. (2023, March 8). The trouble with AI: A conversation with Stuart Russell and Gary Marcus (*Making sense* podcast, Episode 312).

Supplementary readings:

- Bengio, Y. (2023, May 22). How rogue AIs may arise. YoshuaBengio.Org. https://yoshuabengio.org/2023/05/22/how-rogue-ais-may-arise/
- Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, *22*, 71-85.
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press. (Chapter 1, Past developments and present capabilities; Chapter 2, Paths to superintelligence; Chapter 3, Forms of superintelligence)
- Drexler, K.E. (2019). Reframing superintelligence: Comprehensive AI services as general intelligence. Technical Report #2019-1. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf
- Goertzel, B. (2012). Should humanity build a global AI nanny to delay the singularity until its better understood? *Journal of Consciousness Studies*, *19*(1-2), 96-111.
- Heaven, W. D. (2023, June 19). How existential risk became the biggest meme in AI. *MIT Technology Review*. https://www.technologyreview.com/2023/06/19/1075140/how-existential-risk-became-biggest-meme-in-ai/
- Metz, C. (2023, June 10). How could A.I. destroy humanity? *The New York Times*.https://www.nytimes.com/2023/06/10/technology/ai-humanity.html
- Mitchell, M. (2022, December 13). What does it mean to align with human values? *Quanta Magazine*. https://www.quantamagazine.org/what-does-it-mean-to-align-ai-with-human-values-20221213/
- Prinz, J. (2012). Singularity and inevitable doom. *Journal of Consciousness Studies*, *19*(7-8), 77-86.
- Vold, K., & Harris, D. R. (2021). How does artificial intelligence pose an existential risk? In C. Véliz (Ed.), *The Oxford Handbook of digital ethics* (Online edition). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198857815.013.36