PHIL3683
Topics in Applied Philosophy 應用哲學專題

Philosophy and ethics of AI
Course Outline

Time：T 2:30pm-5:15pm	**Location：** LSK 201

Course overview

This course will offer an overview of philosophical, ethical and social issues raised by developments in Artificial Intelligence (AI), from the near-term (e.g. self-driving cars) to the long-term (the rise of machine superintelligence, potentially leading to a technological "Singularity"). After settling some definitional matters, we will consider whether, and if so, how, the many benefits that AI has to offer society can be promoted without generating unacceptable costs. We will also discuss whether AI systems should be recognized as having rights, as well as the possibility of conscious machines.

Advisory to Majors: to be taken in year 2 or above.

Learning outcomes

1. Acquire knowledge of the main issues pertaining to the philosophy and ethics of AI.

2. Have a solid grasp of the relevant philosophical issues.

3. Be able to rigorously articulate and defend a philosophical thesis of their own, in relation to one of the topics covered in the course.

Topics

1. Defining AI
2. AI bias
3. Automation and unemployment
4. Self-driving cars
5. Autonomous weapons
6. AI advisors and moral enhancement
7. Virtual reality and the meaning of life
8. AI companions
9. Do AI & robots have rights?
10. Artificial consciousness
11. The Singularity

Learning activities

One lecture (about 2 hours) per week

In-class interactive activities (about 1 hour per week)

Around 20-30 pages of required readings each week

Assessment scheme

Task nature	***Description***	***Weight***
Attendance & participation	Class attendance and participation in discussion	20%
Oral presentation	Individual or group presentation on one of the required readings	30%
Term paper	Write a paper (min. 3,500 words long, in English) on a topic covered in the course	50%

Grade Descriptor

Please refer to: http://phil.arts.cuhk.edu.hk/~phidept/UG/Grade_descriptors.pdf

Learning resources:

Required:

- Chalmers, D. J. "The Singularity: a Philosophical Analysis". *Journal of Consciousness Studies* 17:9-10 (2010).
- Coeckelbergh, M. "Artificial Companions: Empathy and Vulnerability Mirroring in Human-Robot Relations". *Studies in Ethics, Law and Technology* 4:3 (2010), pp. 1-17.
- Danaher, J. "Virtual Reality and the Meaning of Life". In: Landau, I. (ed.), *Oxford Handbook on Meaning in Life*. Oxford University Press (forthcoming).
- Dubber, M. D., Pasquale, F., & Das, S. (eds.). *The Oxford Handbook of Ethics of AI*. New York: Oxford University Press, 2020.
- Giubilini, A. & Savulescu, J. "The Artificial Moral Advisor: The 'Ideal Observer' Meets Artificial Intelligence". *Philosophy and Technology* 31 (2018), pp. 169-188.
- Liao, S. M. *Ethics of Artificial Intelligence*. New York: Oxford University Press, 2020.
- Müller, V. C. "Ethics of Artificial Intelligence and Robotics". In: Zalta, E. N. (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2021 Edition)*, URL = https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/.
- Nadel, L. (ed.). *Encyclopedia of Cognitive Science*. John Wiley & Sons, 2006.
- Nyholm, S. "The Ethics of Crashes with Self-Driving Cars: a Roadmap", I & II. *Philosophy Compass* 13 (2018), e12507 and e12506.

Recommended:

- Anderson, M. & Anderson, S. L. (eds.). *Machine Ethics*. New York: Cambridge University Press, 2011.
- Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.
- Bramble, B. "The Experience Machine". *Philosophy Compass* 11:3 (2016), pp. 136-145.

- Coeckelbergh, M. "Robot Rights? Towards a Social-Relational Justification of Moral Consideration". *Ethics Inf Technol* 12 (2010), pp. 209-221.
- Coeckelbergh, M. *AI Ethics*. Cambridge, MA: MIT Press, 2020.
- Cole, D.. "The Chinese Room Argument". *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), Zalta, E. N. (ed.), URL = https://plato.stanford.edu/archives/win2020/entries/chinese-room/.
- Danaher, J. "The Philosophical Case for Robot Friendship". *Journal of Posthuman Studies* 3:1 (2019), pp. 5-24.
- Koch, C. & Tononi, G. "Can Machines Be Conscious?". *IEEE Spectrum*, June 2008, pp. 55-59.
- Lara, F. "Why a Virtual Assistant for Moral Enhancement When We Could Have a Socrates?". *Science and Engineering Ethics* 27 (2021), no. 42.
- Searle, J. R. "Minds, Brains, and Programs", *Behavioral and Brain Sciences* 3:3 (1980), pp. 417–424. doi:10.1017/S0140525X00005756.
- Sparrow, R. "Killer Robots", *Journal of Applied Philosophy* 24:1 (2007), pp. 62-77.

Course schedule

Week	***Topics***	***Required reading***	***Teaching Mode***
1	Introduction: defining AI; course overview	Müller, "Ethics of AI & Robotics"	In-campus*
2	AI bias	O'Neil & Gunn, "Near-Term AI & the Ethical Matrix"	
3	Automation and unemployment	James, "Planning for Mass Unemployment"	
4	Self-driving cars	Nyholm, "The Ethics of Crashes with Self-Driving Cars: a Roadmap", parts I & II	
5	Autonomous weapons	Asaro, "Autonomous weapons & the Ethics of AI"	
6	AI advisors and moral enhancement	Giubilini & Savulescu, "The Artificial Moral Advisor"	
7	Virtual reality	Danaher, "Virtual Reality and the Meaning of Life"	
8	AI companions	Coeckelbergh, "Artificial Companions"	
9	Rights for AI/robots?	Basl & Bowen, "AI as a Moral Right-Holder"	
10	Artificial consciousness (1)	Gunderson, "Machine Consciousness"; Searle, "The Chinese Room Argument"	
11	Artificial consciousness (2)	Schneider, "How to Catch an AI Zombie"	
12	The Singularity (1): prospects	Chalmers, "The Singularity" (part one)	
13	The Singularity (2): mind uploading	Chalmers, "The Singularity" (part two)	

* Might change to mix-mode or online in case of unforeseen circumstances

Details of course website

The materials for this course, including readings, lecture slides (PowerPoint), and assignments, will all be

posted on Blackboard as the term unfolds.

Contact details for teacher(s) or TA(s)

Teacher	
Name:	Erler, Alexandre
Office location:	Room 416, Fung king Hey Building
Telephone:	3943 7139
Email:	erleralexandre@cuhk.edu.hk

TA	
Name:	
Office location:	
Telephone:	
Email:	

Academic honesty and plagiarism

Attention is drawn to University policy and regulations on honesty in academic work, and to the disciplinary guidelines and procedures applicable to breaches of such policy and regulations. Details may be found at http://www.cuhk.edu.hk/policy/academichonesty/

With each assignment, students will be required to submit a signed declaration that they are aware of these policies, regulations, guidelines and procedures. For group projects, all students of the same group should be asked to sign the declaration.

For assignments in the form of a computer-generated document that is principally text-based and submitted via VeriGuide, the statement, in the form of a receipt, will be issued by the system upon students' uploading of the soft copy of the assignment. Assignments without the receipt will not be graded by teachers. Only the final version of the assignment should be submitted via VeriGuide.