PHIL 481 Topics in Philosophy: Philosophy of Artificial Intelligence

Term: Fall 2022
Instructor: Professor Jocelyn Maclure
Office: Leacock 930
Email: Jocelyn.maclure@mcgill.ca
Course Schedule: Tues-Thurs 11:35-12:55
Location: Wong 1050
Office hours: Tuesdays 1:30-2:45 and by appointment
Teaching Assistant: Keven Bisson, keven.bisson@mail.mcgill.ca

Course Description

The advances of the past decade in artificial intelligence (AI) have been impressive. From AlphaGo's victory against one the best human Go players to self-driving vehicles, AI is already changing how we think and how we act in all spheres of human life. Progress in computer vision and natural language processing are particularly notable. Computer vision software can be used to identify objects and persons. Decent translations of speech or texts are easily accessible. AI is being used to replace or supplement human judgement in crucial areas such as healthcare, public administration, human resources and the judicial system. Predictive algorithms choose to a large extent the content we are exposed to online and have, in so doing, a powerful influence on our mental life and on our democratic deliberations. After a few decades of stagnation ("AI winters"), the new AI spring is propelled by various types of machine learning algorithms (including "deep learning") and "artificial neural networks". The causes of the AI renaissance and the epistemic strengths and limits of different approaches to machine learning will be reviewed, but no prior technical knowledge in computer science or AI is required for taking this course.

Progress in AI raises a host of complex philosophical questions, both in theoretical and practical philosophy. Our course will straddle both types of question. We will explore fundamental issues such as whether computers can think, have intentional states or be phenomenally conscious. Classic thought experiments such as Alan Turing's imitation game (now called the Turing Test) and John Searle's Chinese Room Argument will be presented and debated. The comparison between animal (human and nonhuman) and machine cognition will be at the forefront of our discussions. A majority of AI researchers and developers think that "artificial general intelligence"(AGI) will be achieved in the coming decades. Current AI systems are narrow; they are good at specific tasks only. Is it plausible to think that an AI will master natural languages, perceive the external world adequately, understand human emotions and other mental states, be capable of moral deliberation, and act competently in their physical environment if they are given an artificial body (robots)? Some philosophers, scientists and technologists even go further by suggesting that the

prospect of “superintelligent” AI systems should be taken seriously. According to theorists such as Nick Bostrom and Stuart Russell, the emergence of artificial superintelligence would create an existential risk for humankind.

In the summer of 2022, an engineer employed by Google opined that LaMDA—a so-called “foundation language model”—was “sentient”. Sentience is usually understood as an entity’s capacity to feel sensations such as pain and pleasure. As such, it appears to require consciousness, i.e. the capacity to have subjective experiences (sensations, emotions, desires, beliefs, etc.). Is it plausible to think that the AI systems are, or will become, conscious? How should we think about the relationship consciousness and intelligence?

These speculative questions are connected to practical ones. How should we think about the moral status of artificial agents capable of acting in the world? Should we see them as the bearers of an intrinsic moral worth and dignity with interests of their own, or rather as artefacts created for fulfilling our needs and interests? Are there lessons to be drawn from the evolution of our ways to treat nonhuman animals?

Moving to applied ethics and political philosophy, the last segment of the course will be devoted to the booming field of “AI ethics”. It is widely known that the decisions made by AI systems can be biased against specific groups, that they lack transparency (the “black box’ or “explainability problem”), that the attribution of moral responsibility for automated decisions is a vexed problem, and that protecting privacy is radically more difficult in the digital age. Moreover, since AI now makes it possible to automate not only manual labor, but also some cognitive tasks, it will have an impact on the distribution of goods such as wealth, jobs, social esteem, and so on. We will see how different theories of justice can help us thinking about the fair distribution of the benefits and risks of automation.

The current hype about AI makes it difficult to assess how transformative it will be. Powerful works of fiction such *Klara and the Sun*, *Machines like me*, *Westworld*, *Her* and *Ex Machina* invite us to think about human life in a world shared which highly intelligent, autonomous and psychologically complex artificial agents. Grand claims about the ongoing cognitive development of AI and about its impacts will be examined with an open mind, but also subjected to a deflationary critique. The hope is that students will be, at the end of the course, in a better position to exercise their own judgment on the impact of AI on human life.

Format

The course will include both lectures and seminar-like discussions in class. The instructor will lecture on various themes in the philosophy of AI whereas the group discussions will focus the reading assignments. There is no textbook; all the readings will be available on MyCourses. The group discussions will start with a team presentation on the required reading. Students must have done the readings and seek to contribute to the group discussion. Guest lecturers may be invited.

Assessments

1) Six commentaries on the reading assignments. Commentaries must be submitted on MyCourses the day before the reading will be discussed in class (so generally on Wednesdays, before midnight). Length: approx. 350 words. 25%

2) Commentary on Karina Vold's lecture. Approx. 350 words. Due date: Friday October 7. 5%

3) One team (2-3 members) presentation. 20-25 mins max. 15%.

4) Summary of the *Taking Stock of AI Ethics* Panel. Approx. 500 words. Due date: Friday November 4. 7.5%

5) Term paper outline: Students must outline the tentative logical structure of their essay and include a briefly annotated bibliography. Due date: Friday November 18. 7.5%

6) Term paper: Students must defend a thesis or position on a philosophy of AI question. Word Limit: 2000 (excluding presentation page and bibliography). Evaluation criteria: (1) understanding of the issue, arguments and literature (17.5 points), (2) argumentative clarity and rigour (17.5 points), (3) bibliographical research and form (5 points). Due Date: Friday December 9. 40%

Late submission of the assignments will be downgraded at a rate of 2 points (not 2%) per day, including weekend/holiday days. Requests for extensions will be considered only when substantiated by a doctor's note or other relevant evidence.

Reading Schedule

	Date	**Reading to do before class**
Week 1	Thursday September 1st	No Reading
Week 2	Tuesday September 6th	
Week 2	Thursday September 8th	Turing, A. M. (1950). Computing machinery and intelligence. *Mind,* 49, 433-460 **Link**: https://www.csee.umbc.edu/courses/471/papers/turing.pdf

	Date	Reading to do before class
Week 3	Tuesday September 13th	
Week 3	ThursdaySeptember 15th	Nagel, T. (1974). What is it Like to Be a Bat?. Philosophical Review,*83*(4), 435-450 **Link** : https://www.jstor.org/stable/2183914?seq=1%252523metadata_info_tab_contents#metadata_info_tab_contents
Week 4	Tuesday September 20th	
Week 4	Thursday September 22nd	**No class.** Replaced by attendance to Karina Vold's lecture on Sept 30, 3:30 pm
Week 5	Tuesday September 27th	Searle, John. R. (1980) Minds, brains, and programs. Behavioral and Brain Sciences, *3*(3): 417-457 **Link** : http://cogprints.org/7150/1/10.1.1.83.5248.pdf
Week 5	ThursdaySeptember 29th	Lecun, Y., Bengio Y., Hinton G. (2015). Deep learning. Nature, 521, 436-444 **Link**: https://www.nature.com/articles/nature14539 Buckner, C. (2019). Deep learning: A philosophical introduction. Philosophy Compass, 14(10), 1-19 **Link**: https://onlinelibrary.wiley.com/doi/10.1111/phc3.12625

	Date	Reading to do before class
Week 6	Tuesday October 4th	
Week 6	Thursday October 6th	Russell, S. (2020). Artificial Intelligence: A Binary Approach. Oxford University Press, Chapter 11 of Ethics of Artificial Intelligence, p. 327-341 **Link**: https://academic.oup.com/book/33540/chapter/287906254
	Tuesday October 11th Thursday October 13th	Fall reading break, no classes
Make up day for Tuesday Week 1	**Friday** October 14th	Liao S. (2020). The Moral Status and Rights of Artificial Intelligence. Oxford university Press, Chapter 17 of Ethics of Artificial Intelligence, 480-504 **Link**: https://academic.oup.com/book/33540/chapter/287907413 Bryson, J. (2010). Robots should be slaves. in Y. Wilks (dir.), et J. Benjamins (chapitre 11, 63-74), *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issue*. **Link :** http://www.cs.bath.ac.uk/~jjb/ftp/Bryson-Slaves-Book09.html

	Date	Reading to do before class
Week 7	Tuesday October 18th	
Week 7	ThursdayOctober 20th	Darling, K. (2012). Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects. We Robot Conference 2012, University of Miami. **Link : https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797**
Week 8	Tuesday October 25th	
Week 8	Thursday October 27th	Wallach. W., Vallor S. (2020). Moral Machines: From Value Alignment to Embodied Virtue, Oxford University Press, chap. 13, 383-412 **Link**: https://academic.oup.com/book/33540/chapter/287906775
Week 9	Tuesday November 1st	
Week 9	ThursdayNovember 3rd	Maclure, J. (2021). AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind. Minds and Machine, 31, 421-438 **Link :** https://link.springer.com/article/10.1007/s11023-021-09570-x

	Date	Reading to do before class
Week 10	Tuesday November 8th	
Week 10	ThursdayNovember 10th	Sax, M. (2018) Privacy from an Ethical Perspective, *The Handbook of Privacy Studies. An Interdisciplinary Introduction,* Bart van der Sloot & Aviva de Groot (ed.), Amsterdam University Press. **Link**: https://www.uva.nl/en/profile/s/a/m.sax/m.sax.html?cb
Week 11	Tuesday November 15th	
Week 11	ThursdayNovembre 17th	Bender, E. et al. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623 **Link**: https://dl.acm.org/doi/10.1145/3442188.3445922 Zoo, J., Schiebinger L. (2018). AI can be sexist and racist— it's time to make it fair. Nature, 559, 324-326 **Link:** https://www.nature.com/articles/d41586-018-05707-8
Week 12	Tuesday November 22nd	

	Date	Reading to do before class
Week 12	ThursdayNovember 24th	James, A. (2020). Planning for Mass Unemployment. Chapter 6 of Ethics of Artificial Intelligence, Oxford University Press, 183-211 **Link :** https://academic.oup.com/book/33540/chapter/287905325
Week 13	Tuesday November 29th	
Week 13	ThursdayDecember 1st	Maclure, J. (2020). The new AI spring: a deflationary view. AI and Society, 35, 747-750 **Link :** https://link.springer.com/article/10.1007/s00146-019-00912-z Skaug Saetra, H et al. (2022). The AI ethicist's dilemma: fighting Big Tech by supporting Big Tech. AI and Ethics, 2, 15-27 **Link**: https://link.springer.com/article/10.1007/s43681-021-00123-7

McGill's policies and recommendations related to COVID-19

This course includes in-person teaching, and learning activities have been planned in accordance with public health directives and McGill's protocols. Please review https://www.mcgill.ca/return-to-campus/

Varia

I tend to think that all electronic devices should be stored away during class, but they are permitted insofar as their use does not disrupt the teaching and learning process. Here is an interesting NPR report on the

subject: https://www.npr.org/2016/04/17/474525392/attention-students-put-your-laptops-away

Please do not record the lectures.

The University requires that the following notices appear on every syllabus:

- McGill University values academic integrity. Therefore, all students must understand the meaning and consequences of cheating, plagiarism and other academic offences under the Code of Student Conduct and Disciplinary Procedures (see www.mcgill.ca/students/srr/honest/ for more information).
- In accord with McGill University's Charter of Students' Rights, students in this course have the right to submit in English or in French any written work that is to be graded.
- In the event of extraordinary circumstances beyond the University's control, the content and/or evaluation scheme in this course is subject to change.