PHIL 481 Topics in Philosophy: Philosophy of Artificial Intelligence

** Note: a revised and updated version will be available before the beginning of the Fall 2022 Term*

Term: Fall 2021
Instructor: Professor Jocelyn Maclure
Office: Leacock 930
Email: Jocelyn.maclure@mcgill.ca
Course Schedule: MW 8:35-9:55
Location: 688 Sherbrooke St, # 355
Office hours: TBA and on 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. In the first half of the semester, 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" 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 in 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.

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 their lessons to be drawn from the evolution of the moral status granted to 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 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. Theories of deliberative democracy can also shed some light on how online platforms, powered by machine learning algorithms, contribute to our current epistemic crisis.

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 status, potential and 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. A few guest lecturers will be invited to present their views.

Assessments

1) Five commentaries on the reading assignments. Commentaries must be submitted on MyCourses the day before the reading will be discussed in class at the latest. Length: 500 words (max). 20% (4 points each)

2) One individual or team (2) presentation. 15 minutes/person max. 15%. Same grade for the two team members.

3) Attendance and participation in the group discussions. 10% : attendance 7 points maximum; participation 3 points maximum. 1 point lost for every absence (max 7 points).

4) Term paper outline: Students must outline the tentative logical structure of their essay and include a briefly annotated bibliography. Due date: November 19th. 10%

5) Term paper: Students must defend a thesis or position on a philosophy of AI question. Word Limit: 3000 (excluding presentation page and bibliography). Evaluation criteria: (1) understanding of the issue, arguments and literature (20 points), (2) argumentative clarity and rigor (20 points), (3) bibliographical research and form (5 points). Due Date: December 10th. 45%

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 justified by exceptional personal circumstances.

Reading Schedule

	Date	Reading to do before class
Week 1	Wednesday September 1st	No Reading
Week 2	Monday September 6th	Labour day, no class

	Date	Reading to do before class
	Wednesday September 8th	Turing, A. M. (1950). Computing machinery and intelligence. *Mind,* 49, 433-460: https://www.csee.umbc.edu/courses/471/papers/turing.pdf
Week 3	Monday September 13th	
	Wednesday September 15th	Nagel, T. (1974). What is it Like to Be a Bat?. Philosophical Review, *83*(4), 435-450 **Link** : https://www-jstor-org.proxy3.library.mcgill.ca/stable/2183914?seq=1#metadata_info_tab_contents
Week 4	Monday September 20th	
	Wednesday September 22nd	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	Monday September 27th	
	Wednesday September 29th	Buckner, C. (2019). Deep learning: A philosophical introduction. Philosophy Compass, 14(10), 1-19 **Link** : https://onlinelibrary.wiley.com/doi/10.1111/phc3.12625
Week 6	Monday October 4th	

	Date	Reading to do before class
	Wednesday October 6th	Russell, S. (2021). Human-Compatible Artificial Intelligence. Stephen Muggleton and Nick Chater (eds.), *Human-Like Machine Intelligence*, Oxford University Press, 1-21 **Link** : https://people.eecs.berkeley.edu/~russell/papers/mi19book-hcai.pdf
Reading break	**Thursday** October 14th (Makeup day for August 30th)	Schwitzgebel, E. et Garza, M. (2018). Designing AI with rights, consciousness, self-respect, and freedom. **Link :** http://www.faculty.ucr.edu/~eschwitz/SchwitzPapers/AIRights2-180604.pdf Bryson, J. (2010). Robots should be slaves. Dans 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
Week 7	Monday October 18th	
	Wednesday October 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	Monday October 25th	

	Date	Reading to do before class
	Wednesday October 27th	Rini, R. (2017). Raising good robots. AEON. **Link :** https://aeon.co/essays/creating-robots-capable-of-moral-reasoning-is-like-parenting Wallach, W., Allen, C. & Smit, I. (2008). Machine morality: Bottom-up and Top-down approaches for modeling human moral faculties. *AI & Society*, *22*(4), 565-582. **Link :** https://www.researchgate.net/publication/220414756_Machine_Morality_Bottom-up_and_Top-down_Approaches_for_Modeling_Human_Moral_Faculties
Week 9	Monday November 1st	
	Wednesday November 3rd	Maclure, J. (2021). AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind. Minds and Machine. **Link :** https://link.springer.com/article/10.1007/s11023-021-09570-x 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
Week 10	Monday November 8th	
	Wednesday November 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

	Date	Reading to do before class
Week 11	Monday November 15th	
	Wednesday Novembre 17th	Binns, R. (2018). What can political philosophy teach us about algorithmic fairness? *IEEE Security & Privacy*, 3, 16, 73-80. **Link :** https://arxiv.org/pdf/1712.03586.pdf 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	Monday November 22nd	
	Wednesday November 24th	James, A. (2020). Planning for Mass Unemployment. Chapter 6 of Ethics of Artificial Intelligence, Oxford University Press, 183-211 **Link :** https://oxford-universitypressscholarship-com.proxy3.library.mcgill.ca/view/10.1093/oso/9780190905033.001.0001/oso-9780190905033-chapter-7
Week 13	Monday November 29th	
	Wednesday December 1st	Maclure, J., Russell, S. (2021). AI for Humanity: The Global Challenge. Lecture Notes in Computer Science, vol. 12. 116-126 **Link :** https://link.springer.com/chapter/10.1007/978-3-030-69128-8_8 Hagendorff, T. (2020). The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. 30, 99–120 **Link :** https://doi.org/10.1007/s11023-020-09517-8

	Date	Reading to do before class

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. It is important, however, to ensure you have read and abide by the following:

- Please review and follow the Health Guidelines for Students, and it is imperative that you understand when to stay home if, for example, you are experiencing COVID-19 symptoms.

- If you develop COVID-19 symptoms while on campus, please follow the required guidelines, which include ensuring you have a mask on, isolate in a closed, private room, immediately call 1-877-644-4545 (Info-Santé) for instructions, and notify the University by calling 514-398-3000.
- **Masks are required in classroom and teaching lab settings**, at all times, and masks will be available for you on campus. Masks are also to be worn when entering and circulating in buildings and classrooms.

- If you are in a situation that might require you to miss some lectures or assignments because of short-term absences due to COVID-19, you are to request an academic accommodation using the online form found under the "Personal" menu in MINERVA; the form is called "**COVID-19 Academic Accommodations Request Form**". You are asked to use this form instead of requesting accommodations directly from your instructor.

Finally, the context of attending University during a pandemic will bring on additional stress and may impact your wellbeing. Please do not hesitate to reach out for support if necessary, and access the many resources available, including Student Services, the Office of the Dean of Students, and your Faculty's Student Affairs Office.

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.