PHIL 680 Problems of Philosophy: Philosophy of Artificial Intelligence

Term: Fall 2023
Instructor: Prof. Jocelyn Maclure
Office: Leacock 930
Email: Jocelyn.maclure@mcgill.ca
Course Schedule: Tues 11:35 am - 2:25 pm
Location: BIRKS 004A
Zoom hours: TBC (https://mcgill.zoom.us/j/7733831777) and by appointment

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 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. Large Language Models and other generative AI took the world by storm and are forcing a major rethink in higher education and in the workplace. AI is being used to replace or supplement human judgement, imagination and expertise in crucial areas such as healthcare, public administration, human resources, the judicial system and the arts. 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 and architectures, including "deep learning" and "artificial neural networks".

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 (commonlly 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 its 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.

Accordingly, they think that answering the “value-alignment problem” or “control problem” is a global priority.

Relatedly, in the summer of 2022, an engineer employed by Google opined that LaMDA—a Large 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 phenomenal consciousness, i.e. the capacity to have subjective experiences. Is it plausible to think that the AI systems are, or will become, conscious? 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 to fulfill 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 second part of the seminar will be devoted to the questions and problems currently discussed in 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 and legal responsibility for an AI system’s decisions and actions 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 participants will be, at the end of the semester, in a better position to exercise their own judgment on the impacts of AI on human life.

Format

The instructor will give short introductory lectures and students will present on the reading assignments. Students must have done the required readings and seek to contribute to group discussions.

myCourses

This seminar has a myCourses site. Assigned readings and course documents can be found there. There is no textbook. All announcements will be posted here, and this is where you’ll turn in your assignments.

Assessment

Participation: You are expected to attend every session, do the assigned readings, and to participate actively and respectfully in each session (by raising questions and/or making comments). 10%

Commentaries/reading responses: 8 short responses to the mandatory readings. These will be due on Mondays by 6:00pm (EST). Length: approx. 350 words. They should raise a thoughtful question about the reading or develop a critical response to an aspect of the reading (or both). Late submissions are not accepted. 20%

One oral presentation on the weekly readings. 20-25 mins. 15%.

Critical Response to a lecture. Due Date November 10th. 5%

Final paper outline: outline of the theme and objective of the paper, its tentative logical structure, and a provisional bibliography. Due Date: November 24th 10%

Final paper: you must defend a thesis or position on a philosophy of AI question. Length: 5000 words (max). Due date: December 15. 40%

Reading Schedule

	Date	**Reading to do before class**
Week 1	September 5th	**Many readings are found in:** Liao, S. Matthew, ed. 2020. *Ethics of Artificial Intelligence*. New York, NY, United States of America: Oxford University Press. **For access, login using McGill Library:** **https://mcgill.on.worldcat.org/oclc/1149361594**
Week 2	September 12th	***Can Machines Think? (1)*** 1/ Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 49, 433-460 **Link:** **https://www.cs.mcgill.ca/~dprecup/courses/AI/Materials/turing1950.pdf**
Week 3	September 19th	***Can Machines Think? (2)*** 2/ Searle, John. R. (1980) Minds, brains, and programs. Behavioral and Brain Sciences, *3*(3): 417-457 **Link: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A**

	Date	Reading to do before class
Week 4	September 26th	***The Moral Status of Advanced AI Systems*** 1/ Chalmers, David J. “Could a Large Language Model Be Conscious?” Boston Review, August 9, 2023. **Link:** https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/ 2/ Schwitzgebel, Eric, and Mara Garza, 'Designing AI with Rights, Consciousness, Self-Respect, and Freedom', in S. Matthew Liao (ed.), *Ethics of Artificial Intelligence* (New York, 2020; online edn, Oxford Academic, 22 Oct. 2020). **Link:** https://doi.org/10.1093/oso/9780190905033.003.0017 *Optional:* 1/ Müller, V.C. Is it time for robot rights? Moral status in artificial entities. *Ethics Inf Technol* 23, 579–587 (2021). **Link:** https://doi.org/10.1007/s10676-021-09596-w 2/ Véliz, C. Moral zombies: why algorithms are not moral agents. *AI & Soc* 36, 487–497 (2021). **Link:** https://doi.org/10.1007/s00146-021-01189-x 3/ Liao, S. Matthew, 'The Moral Status and Rights of Artificial Intelligence', in S. Matthew Liao (ed.), *Ethics of Artificial Intelligence* (New York, 2020; online edn, Oxford Academic, 22 Oct. 2020). **Link:** https://doi.org/10.1093/oso/9780190905033.003.0018 4/ Bryson, Joanna J. (2010). “Robots should be slaves.” In Yorick Wilks (ed.), *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues.* pp. 63-74. **Link:** https://philpapers.org/rec/BRYRSB 5/ 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 5	October 3rd	Guest: Prof. Markus Gabriel, University of Bonn Reading: TBC
	October 10th	No class. Fall Reading Break

	Date	Reading to do before class
Week 6	October 17th	***Artificial General Intelligence/Superintelligence, Existential Risk and the Value-Alignment Problem*** 1/ 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 2/ Bengio, Yoshua. “AI and Catastrophic Risk”, Journal of Democracy, Septermber 2023, https://www.journalofdemocracy.org/ai-and-catastrophic-risk/ *Optional:* 1/ Russell, Stuart J. 2019. *Human Compatible : Artificial Intelligence and the Problem of Control*. New York: Viking. 2/ Bostrom, Nick. 2014. *Superintelligence: paths, dangers, strategies.* Oxford University Press. 3/ Bengio, Yoshua. “FAQ on Catastrophic AI Risks.” yoshuabengio.org. **Link:** https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks/ 4/ 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 7	October 24th	***Deep Learning & Large Language Models*** 1/ Sutton, Richard. “The Bitter Lesson.” incompleteideas.net. March 13, 2019. **Link:** http://www.incompleteideas.net/IncIdeas/BitterLesson.html 2/ Shanahan, Murray. “Talking About Large Language Models.” December 7, 2022. **Link:** https://arxiv.org/abs/2212.03551 3/ Floridi, L. *AI as Agency Without Intelligence*: on ChatGPT, Large Language Models, and Other Generative Models. *Philos. Technol.* 36, 15 (2023). **Link:** https://link.springer.com/article/10.1007/s13347-023-00621-y *Optional:* 1/ Bubeck et al., “Sparks of Artificial General Intelligence; Early experiments with GPT-4.” March 22, 2023. **Link:** https://arxiv.org/abs/2303.12712

	Date	Reading to do before class
		2/ LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* 521, 436–444 (2015). **Link:** https://doi.org/10.1038/nature14539 3/ Buckner, C. (2019). Deep learning: A philosophical introduction. Philosophy Compass, 14(10), 1-19 **Link:** https://onlinelibrary.wiley.com/doi/10.1111/phc3.12625 4/ Bender, E., Gebru, T. 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 5/ Choi, Yejin. (2022). The Curious Case of Commonsense Intelligence. *Daedalus* 151 (2): 139–155. **Link:** https://doi.org/10.1162/daed_a_01906
Week 8	October 31st	***Designing Virtuous Artificial Agents?*** 1/ 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 2/ Rini, Regina. "Creating Robots Capable of Moral Reasoning Is like Parenting: Aeon Essays." Aeon, 2017. **Link:** https://aeon.co/essays/creating-robots-capable-of-moral-reasoning-is-like-parenting
Week 9	November 7th	No class
Week 10	November 14th	***Automation, Distributive Justice and the Meaning of Work*** 1/ Danaher, John (2017). Will Life Be Worth Living in a World Without Work? Technological Unemployment and the Meaning of Life. Science and Engineering Ethics 23 (1):41-64. **Link:** https://philarchive.org/rec/DANWLB 2/ Nieswandt, Katharina (2021). Automation, Basic Income and Merit. In Keith Breen & Jean-Philippe Deranty (eds.), Whither Work? The Politics and Ethics of Contemporary Work. Milton and New York: Routledge. pp. 102–119.

	Date	Reading to do before class
		Link: https://philpapers.org/rec/NIEABI *Optional:* 1/ 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 11	November 21st	***AI's Explainability Problem*** 1/ Buckner, Cameron (forthcoming). "Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour." *British Journal for the Philosophy of Science.* **Link:** https://www.journals.uchicago.edu/doi/10.1086/714960# **Access through McGill Library:** https://mcgill.on.worldcat.org/oclc/41964018 2/ 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 *Optional:* 1/ Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, *32*(4), 661–683. **Link:** https://doi.org/10.1007/s13347-018-0330-6 2/ Vredenburgh, Kate (2021). The Right to Explanation. Journal of Political Philosophy 30 (2):209-229. **Link:** https://onlinelibrary.wiley.com/doi/10.1111/jopp.12262#pane-pcw-references 3/ Vaassen, B. AI, Opacity, and Personal Autonomy. *Philos. Technol.* 35, 88 (2022). **Link:** https://doi.org/10.1007/s13347-022-00577-5
Week 12	November 28th	***Predictive Algorithms, Recommender Systems, Autonomy and Privacy*** 1/ Prunkl, C. (2022). Human Autonomy in the Age of Artificial Intelligence. Nature Machine Intelligence 4 (2):99-101. **Link:** https://philarchive.org/rec/PRUHAI 2/ Yeung, Karen. (2017). "Hypernudge: Big Data as a mode of regulation by design." *Information, Communication & Society*, 20:1, 118-136.

	Date	Reading to do before class
		Link: https://www.researchgate.net/publication/303479231_'Hypernudge'_Big_Data_as_a_mode_of_regulation_by_design 3/ Laitinen, Arto & Sahlgren, Otto (2021). AI Systems and Respect for Human Autonomy. *Frontiers in Artificial Intelligence 4.* **Link:** https://philpapers.org/rec/LAIASA-3 *Optional:* 1/ Jesse, Mathias & Jannach, Dietmar. (2021). "Digital nudging with recommender systems: Survey and future directions." *Computers in Human Behavior Reports* 3. **Link:** https://www.sciencedirect.com/science/article/pii/S245195882030052X 2/ Susser, D., Roessler, B., Nissenbaum, H. (2019). "Online Manipulation: Hidden Influences in a Digital World." 4 *Georgetown Law Technology Review* 1. **Link:** https://philarchive.org/archive/SUSOMHv1
Week 13	December 5th	***Doing AI Ethics*** 1/ Mhlambi, S., Tiribelli, S. Decolonizing "AI Ethics: Relational Autonomy as a Means to Counter AI Harms." *Topoi* 42, 867–880 (2023). **Link:** https://link.springer.com/article/10.1007/s11245-022-09874-2 2/ van Maanen, G. AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics. DISO 1, 9 (2022). **Link:** https://link.springer.com/article/10.1007/s44206-022-00013-3 *Optional:* 1/ Birhane A. (2021). "Algorithmic injustice: a relational ethics approach." *Patterns* 2(2):100205. **Link:** https://doi.org/10.1016/j.patter.2021.100205 2/ Russo, F., Schliesser, E., Wagemans, J. (forthcoming). "Connecting ethics and epistemology of AI." *AI and Society*: 1-19. **Link:** https://link.springer.com/article/10.1007/s00146-022-01617-6 3/ Cole M., Cant C., Ustek Spilda F., Graham M. (2022). "Politics by Automatic Means? A Critique of Artificial Intelligence Ethics at Work." *Frontiers in Artificial Intelligence 5.* **Link:** https://www.frontiersin.org/articles/10.3389/frai.2022.869114/full

	Date	Reading to do before class
		4/ Seger, E. (2022). “In Defence of Principlism in AI Ethics and Governance.” *Philos. Technol.* 35. **Link:** https://doi.org/10.1007/s13347-022-00538-y

Varia

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).
- McGill’s Teaching and Learning Services’ Recommendations on Generative AI: https://www.mcgill.ca/tls/channels/news/stl-approves-recommendations-generative-ai-349064
- 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.