



PHIL 481-001 Topics in Philosophy: Artificial Intelligence

Instructor: Daniel Harris
Email: daniel.harris2@mail.mcgill.ca
Location: BIRKS 111
Days & Time: Tues/Thurs 2:35-3:55

Course Description: Research in the field of Artificial Intelligence (AI) is developing at a rapid pace. AI systems sort your email, diagnose diseases, locate new galaxies, recommend movies, drive cars and have defeated our best minds in chess and go. As this technology continues to mature we should expect it to embed itself more and more in the fabric of our lives. With these changes come tremendous opportunities for human advancement, including those in medicine and health, education, transportation, science, environmental sustainability, and economic growth. Equally, though, such changes represent substantial risks, such as the possibility of wide-spread labor displacement, increased socio-economic inequity, oligopolistic market structures, totalitarianism, and, in extreme circumstances, the subjugation or extinction of humanity. This course will cover the philosophical issues that emerge out of the development of current and future AI systems. It is divided into three sections: (1) *classic philosophical challenges to the artificial general intelligence enterprise*, (2) *AI ethics & politics*, and (3) *AI safety & existential risk*.

Course Objectives: At the completion of this course students will be able to:

- Demonstrate familiarity with three core debates in the history of the philosophy of AI.
- Explain the philosophical issues surrounding the moral status of AI.
- Understand the ethical and socio-political obstacles involved in designing AI systems.
- Exhibit knowledge of the problems surrounding AI safety and existential risk mitigation.

Assessment:

- 10% Attendance & Participation.
- 20% Reading Assignments.
- 25% Term Paper Outline.
- 45% Term Paper.

Reading Assignments: Students will complete two assignments on the weekly readings. Each set will be worth 10%, for a total of 20% of your final grade. Assignments should include (i) a well thought out discussion question, and (ii) a critical engagement with at least one position advanced in the reading(s).

Term Paper Outline: Students will submit an in-depth outline of their term paper detailing (i) the argumentative structure of their essay, and (ii) a brief overview of the literature they will be engaging with. Due date: March 14th.

Term Paper: Students will complete an end of term essay that addresses some aspect of the course material and readings. Due date: April 12th. Word Limit: 3500-4000.

Course Materials: All course readings will be available through myCourses.

Topics & Readings

(Readings will be determined on a week-by-week basis, and will be announced on myCourses.)

Artificial General Intelligence (AGI): A Philosophical Engagement

(1) The Turing Test

In his 1950 article *Computing Machinery and Intelligence* Alan Turing predicted that at the end of the century “the use of words and general educated opinion will have altered so much that one will be able to speak of machines as thinking without expecting to be contradicted”. It is certainly true that in the years that have followed the notion of a thinking machine has entered the public lexicon. Yet despite opinions having shifted radically as a result of Turing’s insights, the possibility of endowing a machine with general intelligence still remains very much an open question. In this section we will critically engage with Turing’s seminal paper via the work of John Searle, and take up the debate surrounding artificial general intelligence these two authors initiated.

An Operational Definition of Intelligence & The Turing Test

- Alan M. Turing (1950). Computing Machinery and Intelligence. *Mind* 59, pp.433-60.

Searle & The Chinese Room Thought Experiment

- John R. Searle (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences* 3(3), pp. 417-457.
- Josef Moural (2003). The Chinese Room Argument. In Barry Smith (ed.), John Searle. Cambridge: Cambridge University Press. Excerpt pp. 214-226.

Responding to Searle: The Systems Reply

- Margaret A. Boden (1988). Escaping From The Chinese Room. In John Heil (ed.), *Computer Models of Mind*. Cambridge University Press, pp. 253-266.
- Suggested Reading: Jack B. Copeland (1993). The Curious Case of the Chinese Gym. *Synthese* 95(2), Excerpt pp. 173-177.
- Suggested Reading: David Anderson & Jack B. Copeland (2002). Artificial Life and the Chinese Room Argument. *Artificial Life*, 8(4), Excerpt pp. 8-11.

(2) Is Everything A Computer?

A widely held assumption of the artificial general intelligence enterprise is that the right kind of computational structures would suffice for the possession of a mind; a belief that is rooted in the view that the human brain is a kind of digital computer. In this section we will critically engage with this thesis via the work of John Searle and his twofold argument that (1) nothing is intrinsically computational, and that (2) everything, at some level of description, can be ascribed a computational assignment.

Social Ontology & The Observer-Relativity of Computation

- John R. Searle (1990). Is The Brain a Digital Computer? *Proceedings and Addresses of the American Philosophical Association* 64(3), pp. 21-37.
- John Searle (1995). The Construction of Social Reality. Free Press, Excerpt, pp.5–29.
- Roger Fellows (1995). Welcome to Wales: Searle on the Computational Theory of Mind. *Royal Institute of Philosophy Supplement* 38, pp. 85-97.

Responding to Searle: Implementation vs. Implimentational Capacity

- Ned Block (2003). The Mind as Software in the Brain. In J. Heil (ed.), *Philosophy of Mind: A Guide and Anthology*. Oxford University Press, Excerpt pp. 15–16.

(3) The Frame Problem

According to Clark Glymour, instances of the epistemological frame problem are of the form: given an enormous amount of stuff, and some task to be done using some of the stuff, what is the relevant stuff for the task? Humans are exceptionally good at solving this problem, but what about AI systems? In this section we will explore this question via an engagement with the work of Daniel Dennett and John Searle.

The Epistemological Frame Problem

- Daniel Dennett (1984). Cognitive wheels: The Frame Problem of AI. In Christopher Hookway (ed.), *Minds, Machines and Evolution*. Cambridge University Press.
- Murray Shanahan (2016) The Frame Problem. The Stanford Encyclopedia of Philosophy.

Unarticulated Background Practices & The Epistemological Frame Problem

- John Searle (1980). The Background of Meaning. In J. Searle, F. Kiefer & M. Berwisch (eds.) *Speech Act Theory and Pragmatics*. Dordrecht, pp. 221–233.

AI Ethics & Politics

(4) Asimov’s Three Laws of Robotics & The Moral Status of AI

In this section we will use Asimov’s “Bicentennial Man” to discuss a number of metaethical issues concerning machine ethics and the moral status of AI.

- Isaac Asimov (1976). The Bicentennial Man. In *The Bicentennial Man and Other Stories*. Doubleday, pp. 138–172.
- Susan Leigh Anderson (2011). The Unacceptability of Asimov’s Three Laws of Robotics as a Basis for Machine Ethics. In M. Anderson & S. Anderson (eds), *Machine Ethics*. Cambridge University Press, pp. 285. 296.

(5) The Moral Status of AI (Cont.)

Moral Status, as a concept, is a vexed philosophical issue. On Mary Anne Warren’s view “[t]o have moral status is to be morally considerable [...], it is to be an entity towards which moral agents have, or can have, moral obligations”. In this section we will use this conceptual framework to take up the question: *should we ascribe moral status to AI systems?*

- Mary Anne Warren (1997). *Moral Status: Obligations to Persons and Other Living Things*. Clarendon Press, Excerpt pp. 4–17.
- John P. Sullins (2006). When is a Robot a Moral Agent? *International Review of Information Ethics* 6 (12), pp.23-30.
- Michael LaBossiere (2017). Testing the Moral Status of Artificial Beings; or I’m Going to Ask You Some Questions”. In P. Lin, K. Abney & R. Jenkins (eds.) *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Oxford University Press, pp. 1–22.
- Deborah G. Johnson (2006). Computer systems: Moral Entities But Not Moral Agents. *Ethics and Information Technology* 8(4), pp. 195–204.

(6) The Design of Ethical AI Systems

The field of artificial intelligence is moving ever closer to the creation of fully autonomous agents. As this research continues to development the question of how AI systems will make moral decisions, and the ethical principles they should deploy in doing so, becomes increasingly important. In this section we will explore the theoretical challenges that the design and implementation of a moral autonomous agent poses.

- Colin Allen, Gary Varner & Jason Zinser (2000). Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental & Theoretical Artificial Intelligence*. 12, pp. 251–261.
- Edmond Awad *et al.* (2018). The Moral Machine Experiment. *Nature*. 563, pp. 59–64.
- Roman Yampolskiy & Joshua Fox (2013). Safety Engineering for Artificial General Intelligence. *Topoi*. 32(2) pp. 217–226.

(7) The AI Arms Race

Nation states and industry are increasingly framing the discussion around the development of artificial intelligence as a race for strategic advantage. As these groups compete for a limited pool of researchers, and announce ambitious strategic plans aimed at establishing AI research leadership, the question of how this rhetoric will effect the current socio-political landscape is pressing. In this section we will assess the potential risks an AI arms race poses, and the steps we might take to mitigate them.

- Stephen Cave & Seán ÓhÉigeartaigh (2018). An AI Race for Strategic Advantage: Rhetoric and Risks. *AAAI/ACM Conference on AI, Ethics, and Society*.
- Edward Moore Geist (2016). It’s Already Too Late to Stop the AI Arms Race — We Must Manage it Instead. *Bulletin of The Atomic Scientists*. Vol. 72, No. 5, pp. 318–321.

(8) Promoting Beneficial AI Research & Development

In recent years there has been a move towards the promotion of the development of AI that is safe and beneficial to society. In this section we will look at the social challenges involved in promoting this kind of design methodology, and assess whether it is both a viable, and politically realistic, research agenda.

- Seth D. Baum (2017). On the Promotion of Safe and Socially Beneficial Artificial Intelligence. *AI and Society*, 32(4), pp. 543–551.
- Nick Bostrom (2017). Strategic Implications of Openness in AI Development. *Global Policy* 8 no. 2, pp. 135–148.

(9) Narrow AI & Social Robots

The emergence of autonomous robots which are designed to interact with humans on a social level pose a number of philosophical issues, not least of which is the position they should occupy in our conceptual, physical, economic, legal, and moral world. In this section we will take up these and related issues surrounding social robotics.

- Kate Darling (2016). Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects. In R. Calo, M. Froomkin, & I. Kerr (eds.), *Robot Law*. Edward Elger, pp. 213-233.
- Deborah G. Johnson & Mario Verdicchio (2018). Why Robots Should Not Be Treated Like Animals. *Ethics and Information Technology*, 20(4) pp. 291–301.

Superintelligence, AI Safety, & Existential Risk

(10) Superintelligence & The “Singularity”

The “singularity” describes a process whereby the event of humanity creating a machine that is more intelligent than itself leads to an explosion of ever-greater levels of intelligence as each generation of a machine creates, in turn, a more intelligent iteration. The possibility of the “singularity” raises a number of philosophical and practical issues which we will take up via an engagement with David Chalmers’ influential paper *The Singularity: A Philosophical Analysis*.

- David Chalmers (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*, 17(9-10) Excerpt, pp. 1–15 & 19–56.
- Ben Goertzel (2012). Should Humanity Build a Global AI Nanny to Delay the Singularity Until Its Better Understood? *Journal of Consciousness Studies*, 19, No. 1-2, pp. 96–111

(11) The “Singularity” & The Computer Simulation Hypothesis

In the previous section we looked at Chalmers’ argument for the inevitability of the “singularity”; that is, an explosion of intelligent machines that will eventuate in the creation of a superintelligence. In this section we will look at an argument which suggests that if Chalmers is right then (1) it is probable that the “singularity” has already happened, and (2) that this entails that we are likely living inside a computer simulation created by a superintelligence.

- Jess Prinz (2012). Singularity and Inevitable Doom. *Journal of Consciousness Studies*, 19 (7-8):77–86.
- Suggested Reading: Nick Bostrom (2003). Are We Living in a Computer Simulation? *Philosophical Quarterly*, 53 (211), pp. 243–255.

(12) Artificial General Intelligence (AGI) & Existential Risk Analysis

What sorts of prediction can we make regarding human-AGI interaction and, of those predictions, can we identify certain scenarios that represent an existential risk to humanity? In this section we will engage with this question by looking at both the Orthogonality Thesis and the Instrumental Convergence Thesis as advanced by Nick Bostrom and Stephan M. Omohundro.

- Nick Bostrom (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, 22(2), pp. 71–85.
- Stephan M. Omohundro (2008). The Basic AI drives. In P. Wang, B. Goertzel & S. Franklin (eds.) *Proceedings of the AGI08 Workshop*. IOS Press, pp. 483–492.
- Stuart Armstrong (2013). General Purpose Intelligence: Arguing the Orthogonality Thesis. *Analysis and Metaphysics*, 12, pp. 68–84.

Responding to Bostrom & Omohundro

- Ben Goertzel (2015). Superintelligence: Fears, Promises and Potentials. *Journal of Evolution and Technology*, Vol. 24 Issue 2. Excerpt pp. 63–69
- Olle Häggström (2018). Challenges to the Omohundro–Bostrom Framework for AI Motivations. *Foresight*, pp. 1–16.

Course Policies

McGill Policies & Statements

Language of Submission: 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. This does not apply to courses in which acquiring proficiency in a language is one of the objectives. Note: In courses in which acquiring proficiency in a language is one of the objectives, the assessments shall be in the language of the course.

Academic Integrity: 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).

General Policies

As the instructor of this course I endeavor to provide an inclusive learning environment. However, if you experience barriers to learning in this course, do not hesitate to discuss them with me and the Office for Students with Disabilities, 514-398-6009.

Mobile computing and communications devices are permitted in class but students are encouraged to limit their use during lectures. The recording of lectures are not permitted without expressed permission from the instructor.

Late work will not be accepted unless there is a serious excuse and corresponding documentation is provided.