PHIL 481-001 Topics in Philosophy: Arti cial 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 eld of Arti cial 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 arti cial 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

Topics & Readings

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

Arti cial 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

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 Arti cial 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 Arti cial 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 eld of arti cial intelligence is moving ever closer to the creation of fully autonomous

Seth D. Baum (2017). On the Promotion of Safe and Socially Bene cial Arti cial 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 E ects 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' in uential 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*, 1 9 (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