York University
Philosophy of Artificial Intelligence
AP/PHIL/COGS 3750 3.00 (Lect 01)
Winter 2021

Course Type: Lecture | Thursdays, 2:30pm (EST), 3 Hours | Location: Zoom | Cat# M73K01 (AP COGS) / W55M01 (AP PHIL)

Important Dates: Jan 11 (start of term), Jan 14 (first class), Feb 13—19 (winter reading week), Mar 12 (last day to drop without receiving a grade), Apr 8 (last class), Apr 12 (winter classes end), April 13 (last day to submit work for the term), April 14—28 (winter exam period)

Course Instructor: Michael Barkasi (barkasi@yorku.ca)
Office Hours: by Zoom, Thursdays, 1:30-2:30pm (EST); alternative times possible. Appointment Required (please email).

Prerequisites: One of AP/PHIL/COGS 2160 3.00 or AP/PHIL 2240 3.00

Technical requirements for taking the course: eClass access and Zoom. Students are *strongly* encouraged to attend Zoom lectures on Thursdays and actively participate with mic and video, but this is not required. (Students attending the Zoom meeting may leave their cams off and mics muted, if they wish.) Lectures will be recorded and made available through eClass, for those who cannot attend. (Discussion time with students will not be recorded, so students not attending the Zoom meeting will miss this aspect of class.)

Here are some useful links for student computing information, resources and help:
Student Guide to Moodle | Zoom@YorkU Best Practices | Zoom@YorkU User Reference Guide | Computing for Students Website | Student Guide to eLearning at York University

Times and locations: This is a remotely delivered course. There will be lectures and discussion by Zoom during the scheduled 2:30-5:30pm (EST) time slot each Thursday. A link for a reoccurring Zoom meeting will be posted to eClass, along with recordings of lectures (but not discussion) for those who cannot attend live. Although attendance for the normal Zoom meetings is not required, you will be required to take your midterm exam during the scheduled Thursday 2:30-5:30pm (EST) time slot for week 8 (March 4); you'll also be required to take the final exam during the time slot assigned to the course during the end-of-semester exam period. Please note that this is a course that depends on remote teaching and learning. There will be no in-person interactions or activities on campus.

Virtual office hours: by Zoom, Thursdays, 1:30-2:30pm (EST), or at a time we mutually agree on. Appointment required in either case (please email to setup an appointment and for the Zoom link). Please do not hesitate to contact me (the course director) by email if you have questions, comments, or concerns.

Required Course Text / Readings: All required reading will be available either freely on the internet, or through a York library link. Links will be posted to eClass. There is no textbook to purchase.

Expanded Course Description: An introduction to philosophical issues in Artificial Intelligence (AI). The goal is for students to be able to gain a basic understanding of the cognitive architectures used by Al programmers, and reflect critically on research in Al from a philosophical perspective. We'll look at challenges to the idea that AI is "real" intelligence, ask whether AI can be creative, and discuss whether AI can be conscious. We'll also look at the ethical dimensions of AI. Can AI systems be moral agents with rights and responsibilities? What are the ethical implications of AI-powered applications like directed marketing, facial recognition, and criminal-sentencing evaluation? Will AI usher in a technological utopia, or a dystopian nightmare?

Organization of the course: Live Zoom lecture and discussion during the scheduled meeting time, i.e. Thursdays, 2:30-5:30pm (EST). The three-hour block will be broken up into 15- to 30-minute lectures with breaks and discussion time in between. Attendance is strongly encouraged, but not mandatory. Recordings of lectures (but not discussion periods) will be made available on eClass after each meeting. There is also a (semi)weekly quiz which students must take on eClass. The quiz opens at the end of each meeting and must be completed by the start of the next meeting. There is a 20-minute time limit on the quiz and students can take it any time during the week it's open on eClass. There will be no quiz the week of the midterm and the weeks the paper/paper revision is due (see the below schedule). A midterm exam will be given during our scheduled meeting time in week 8 (March 4). You will be required to take the exam during the scheduled class time that week. Similarly, you'll also be required to take the final exam during the time slot assigned to the course during the end-of-semester exam period. There is a term paper due by the start of class week 11, March 25, with an optional revision (replacing the grade of the original term paper) due by the start of class week 13, April 8.

Course Learning Objectives: (1) Understand the rudiments of AI architectures. (2) Understand philosophical questions about the intelligence, creativity, rationality, personhood, moral standing, and consciousness of AI. (3) Learn how to engage in sustained philosophical analysis through the medium of short essays (midterm and final exam) and a term paper. (4) Become a more reflective user and/or designer of AI systems, so that you have more awareness and control of how these systems shape your life and the life of others.

Evaluation:

- **Weekly Quizzes** (25% of total grade): At the end of most classes (see the schedule) a quiz will open covering the material from that class. It must be completed by the start of the next class. The quizzes will be multiple-choice, 5-10 questions. The quizzes are open-book and unproctored. The quizzes will be taken on eClass and are timed (20 minutes). No extensions or redos will be given, even for technical difficulties. Instead, the lowest two quiz scores will be dropped.

- Midterm Exam (20% of total grade): There will be a midterm exam given in lieu of our week 8 meeting (March 4). It will consist of short-essay questions. Students are expected to take their midterm exam, over eClass, during the normal three-hour class time scheduled for that week. The midterm is open-book and unproctored.
- Final exam (20% of total grade): There will be a final exam given during a time scheduled during the exam period (April 14—28, time/date TBA). Like the midterm, it will be held over a 3 hour window, given on eClass, and consist of short-essay questions (plus some survey and reflection questions). The final is open-book and unproctored.
- Paper (35% of total grade): 2-3,000 words on a topic of your choice, with instructor approval. Example topics will be provided for those who need suggestions. A rubric will be provided and expectations clearly laid out, no later than the midterm. Papers are due by the start of class, week 11 (March 25). They will be submitted through eClass; Turnitin will be used. An optional revision, based on instructor feedback, will be due by the start of our last class meeting. The grade on the revised paper will replace the original paper grade (if it's higher).

Weighting of Course: Weekly quizzes: 25% | Midterm: 20% | Term paper: 35% | Final exam: 20%

Course policies:

- Missed quizzes: No extensions or redos will be given for quizzes, no exceptions, including for technical difficulties.
- Midterm makeups: Alternative arrangements for the midterm will be granted for students who are unable to take the midterm at the scheduled time, if there are circumstances that necessitate it. These alternative arrangements must be made ahead of time (preferably at least a week in advance); students who miss the midterm without alternative arrangements in place ahead of time will be allowed to makeup the midterm at the instructor's discretion, but a legitimate excuse is required. If you miss the midterm without a legitimate excuse, you may be allowed to makeup the midterm, but with a point reduction worth 10% of the total possible points (e.g., if the midterm is scored on a scale of 30 points, the reduction will be 3 points).
- Final exam makeups: Alternative arrangements and makeups will not be allowed, except as required by university policy.
- Open-book policy: It is expected that students will complete quizzes, midterm, and final exam on their own without help from any other person. But, students may use their notes, the articles covered in class, and other materials that have been provided over eClass. Google or other internet searching is not allowed, nor, once you have started the quiz or exam, are you allowed to ask a classmate or friend for help in any way. If you take a quiz or exam before a classmate, you cannot communicate to them the questions or anything that may give them an unfair advantage. "Open-book" means only that you are allowed to use your notes, articles covered in class, and other materials provided over eClass. When in doubt, ask for clarification.
- Paper: Late papers will be accepted, with a possible reduction in points, although the instructor reserves the right to refuse accepting a late paper if the circumstances are unreasonable or the paper is extremely late. Papers turned in late without reasonable mitigating circumstances or prior arrangements may incur a reduction in points worth up

to 10% of the total possible points (e.g., if the paper is scored out of 30 points, the reduction may be up to 3 points). It's best that you contact me as soon as possible if you're going to be late (preferably before the due date).

- **Turnitin:** To promote academic integrity in this course, students will be normally required to submit their written assignments to Turnitin (via eClass) for a review of textual similarity and the detection of possible plagiarism. In so doing, students will allow their material to be included as source documents in the Turnitin.com reference database, where they will be used only for the purpose of detecting plagiarism. The terms that apply to the University's use of the Turnitin service are described on the Turnitin.com website.
- **Attendance:** As noted above, it's not required that you attend the live Zoom classes. While these classes are your best opportunity to interact with me and ask questions about the material, nonattendance will not be held against you in any way. No preference is given to students who attend the live Zoom classes.
- **Zoom sessions:** For those attending the live Zoom classes, you are not allowed to take any screenshots or recordings of any kind. This is to respect the privacy of your fellow students. Recordings of the lecture portions (which only involve myself) and my slides will be made available on eClass. You also do not have permission to reproduce any lecture recordings on any platforms or websites outside of eClass.
- **Academic honesty and integrity:** In this course, we strive to maintain academic integrity to the highest extent possible. Please familiarize yourself with the meaning of academic integrity by completing SPARK's Academic Integrity module at the beginning of the course. Breaches of academic integrity range from cheating to plagiarism (i.e., the improper crediting of another's work, the representation of another's ideas as your own, etc.). All instances of academic dishonesty in this course will be reported to the appropriate university authorities, and can be punishable according to the Senate Policy on Academic Honesty.
- **Intellectual property:** All course material (this syllabus, power points, assignments, paper rubrics, etc), except the outside assigned articles, is the intellectual property of the course instructor and cannot be reproduced in any way without my permission. Assigned articles are the intellectual property of their respective copyright holders and usually cannot be reproduced or posted publicly.
- **Student conduct:** All students are expected to treat their fellow students and the instructor with respect and charity. Especially through mediums like Zoom and the course eClass, no form of harassment, trolling, or disrespect will be tolerated.
- **Further links:** For more information on relevant university policies, please see: Student Rights & Responsibilities, Academic Accommodation for Students with Disabilities, and Important Course Information.

Special Accommodations: The course instructor is committed to fairly accommodating students with disabilities. Please contact me and Student Accessibility Services (https://accessibility.students.yorku.ca) as soon as possible, and we will all work together to find a fair accommodation.

Schedule:

This schedule is tentative. Assigned readings are listed on the next page and will be posted directly to eClass. The readings assigned for each week need not be read before the start of class, but should be read before you take the quiz for that week's material. Quizzes may include questions from the essential readings not covered in lecture.

Wk1: Jan 14	Introduction to AI: History, Architectures, and Examples	
Wk2: Jan 21	Classical challenges: Physical symbol systems, the Turing test, and the Chinese room	Quiz 1 must be taken by the start of class
Wk3: Jan 28	Embodied AI: Why be a robot?	Quiz 2 must be taken by the start of class
Wk4: Feb 4	Can AI be creative? Gödel's incompleteness theorem, Art, and AlphaGo	Quiz 3 must be taken by the start of class
Wk5: Feb 11	What is general artificial intelligence?	Quiz 4 must be taken by the start of class
Wk6: Feb 18	No Class: Reading Week	
Wk7: Feb 25	Belief update: The frame problem, isotropy, and Quinean holism	Quiz 5 must be taken by the start of class
Wk8: Mar 4	Midterm Exam	
Wk9: Mar 11	Phenomenal consciousness: Is there something it's like to be an AI?	Quiz 6 must be taken by the start of class
Wk10: Mar 18	The normative status of AI: Rationality, rights, personhood, and moral standing	Quiz 7 must be taken by the start of class
Wk11: Mar 25	The practical ethics of AI: Facial recognition, Facebook advertising, and criminal-sentencing decisions	Term Paper due by start of class
Wk12: Apr 1	AI dystopias: Existential risk and algorithmic human engineering	Quiz 8 must be taken by the start of class
Wk13: Apr 8	AI utopias: From Siri to protein-folding to digital afterlife	Term Paper revision due by start of class
Final Exam Period: Apr 14—28	Date/Time TBA	

Readings:

Readings for each week do not need to be done before the start of the week's class. Most of what you need for the quizzes will be covered in class. Quizzes may have a question or two not directly covered in class, but covered in the essential readings. Even for those questions covered in class, class alone may not afford you with a deep enough understanding to answer the question (some questions will be nuanced and difficult), and so the readings are likely to help you understand the material at the level needed for the week's quiz.

Readings fall into three categories: Essential, Supplemental, and Related-interesting. Essential readings are selected to fill a very specific learning objective of the week. If you do not read them, you will not learn everything offered in this class. Supplemental readings provide extra background you may find helpful in filling out your understanding. Related-interesting readings aren't necessary for the class (and we may not cover them at all), but cover interesting material you may find worth reading. Please keep an eye on eClass, as I may occasionally add a reading, remove a reading, or switch a supplemental/related-interesting reading to essential!

Note: You might find *The Cambridge Handbook of Artificial Intelligence* a useful supplement to this class. We'll use it as a textbook, of sorts. I've listed it often as a supplemental reading (under the abbreviation "CHAI"). You can access the entire book, for free, through your passport York account via this link:
https://ocul-yor.primo.exlibrisgroup.com/permalink/01OCUL_YOR/q36jf8/alma991036301510005164

Links for the rest of the readings will be posted to eClass.

- Wk 1 (Jan 14): Introduction to AI: History, Architectures, and Examples
  - Essential Readings: CHAI chapter 2
  - Supplemental Readings: CHAI chapters 1, 4, 5
- Wk 2 (Jan 21): Classical challenges: Physical symbol systems, the Turing test, and the Chinese room
  - Essential Readings: Turing (1950) "Computing machinery and intelligence",
    Searle (1980) "Minds, brains and programs" (pp. 417—424),
    Lacker (2020) "Giving GPT-3 a Turing Test",
    Ouimet (2020) "Coronavirus (COVID-19)"
  - Supplemental Readings: CHAI chapters 3, 10,
    Alammar (2018) "The illustrated transformer"
    Alammar (2019) "The illustrated GPT-2 (Visualizing transformer language models)",
    Alammar (2020) "How GPT3 Works – Visualizations and animations"
  - Related Interesting Readings: Bringsjord, Bello, & Ferrucci (2001) "Creativity, the Turing test, and the (better) Lovelace test",
    Weinberg (2020) "Philosophers on GPT-3 (updated with replies by GPT-3)"
- Wk 3 (Jan 28): Embodied AI: Why be a robot?
  - Essential Readings: Brooks (1991) "Intelligence without representation",

Dreyfus (1992) *What Computers Still Can't Do: A Critique of Artificial Reason*, chapter 7 (pp. 235—255)
    - Supplemental Readings: CHAI chapters 6, 13
    - Related Interesting Readings: Wolfendale (2020) "Artificial bodies and the promise of abstraction"
- Wk 4 (Feb 4): Can AI be creative? Gödel's incompleteness theorem, Art, and AlphaGo
    - Essential Readings: Lucas (1961) "Minds, Machines and Gödel", Boden (2009) "Computer models of creativity"
    - Supplemental Readings: CHAI chapters 3, 7, 8, Piccinini (2003) "Alan Turing and the mathematical objection"
    - Related Interesting Readings: Coeckelbergh (2017) "Can machines create art?"
- Wk 5 (Feb 11): What is general artificial intelligence?
    - Supplemental Readings: CHAI chapters 7, 8
    - Related Interesting Readings: Chalmers (2020) "GPT-3 and general intelligence", Wang and Goertzel (2012) *Theoretical Foundations of Artificial General Intelligence*, Laird et al. (1987) "SOAR: An architecture for general intelligence", Rosenbloom et al. (1991) "A preliminary analysis of the Soar architecture as a basis for general intelligence"
- Wk 6 (Feb 18): No Class: Reading Week
- Wk 7 (Feb 25): Belief update: The frame problem, isotropy, and Quinean holism
    - Essential Readings: Quine (1951) "Two dogmas of empiricism" (pp. 39—43), Fodor (1983) *The Modularity of Mind*, chapter 4 "Central systems"
    - Supplemental Readings: CHAI chapters 9, 11
- Wk 8 (Mar 4): Midterm Exam
- Wk 9 (Mar 11): Phenomenal consciousness: Is there something it's like to be an AI?
    - Essential Readings: Chalmers (1996) *The Conscious Mind: In Search of a Fundamental Theory*, chapters 7 (pp. 247—75), 9 (pp. 313—32).
    - Related Interesting Readings: CHAI chapter 12, Bechtel (1995) "Consciousness: Perspectives from symbolic and connectionist AI", Bringsjord (2007) "Offer: On billion dollars for a conscious robot; if you're honest, you must decline", Molyneux (2012) "How the problem of consciousness could emerge in robots", Manzotti and Chella (2018) "Good old-fashioned artificial consciousness and the intermediate level fallacy"
- Wk 10 (Mar 18): The normative status of AI: Rationality, rights, personhood, and moral standing
    - Essential Readings: Bryson (2010) "Robots should be slaves", Levy (2009) "The ethical treatment of artificially conscious robots", Marko (2019) "Robot rights – a legal necessity or ethical absurdity?"
    - Supplemental Readings: CHAI chapters 11, 14, 15
    - Related Interesting Readings: Rini (2017) "Raising good robots", Coeckelbergh (2010) "Robot rights? Towards a social-relational justification of moral consideration"

- Wk 11 (Mar 25): The practical ethics of AI: Facial recognition, Facebook advertising, and criminal-sentencing decisions
    - Essential Readings: Hao (2019) "AI is sending people to jail—and getting it wrong",
      Hao (2020) "The coming war on the hidden algorithms that trap people in poverty",
      Tufekci (2015) "Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency",
      Mittelstadt et al. (2016) "The ethics of algorithms: Mapping the debate"
    - Supplemental Readings: CHAI chapters 7, 8, 15
    - Related Interesting Readings: Watney (2017) "It's time for our justice system to embrace artificial intelligence",
      Sharkey (2020) "Can we program or train robots to be good?"
- Wk 12 (Apr 1): AI dystopias: Existential risk and algorithmic human engineering
    - Essential Readings: Barkasi (2020) "Algorithmic human engineering",
      Leslie (2021) "Are we the cows of the future?",
      Turchin and Denkenberger (2020) "Classification of global catastrophic risks connected with artificial intelligence",
    - Supplemental Readings: CHAI chapter 15,
      Frischmann (2020, podcast) "Re-engineering humanity (part one)",
      Bostom (2002) "Existential risks",
      Bostrom (2009) "The future of humanity"
    - Related Interesting Readings: Danaher (2017) "Will life be worth living in a world without work? Technological unemployment and the meaning of life",
      Loi (2015) "Technological unemployment and human disenhancement",
      Bostrom (2012) "The superintelligent will: Motivation and instrumental rationality in advanced artificial agents"
- Wk 13 (Apr 8): AI utopias: From Siri to protein-folding to digital afterlife
    - Essential Readings: Graziano (2016) "Why you should believe in the digital afterlife",
      Callaway (2020) "'It will change everything': AI makes gigantic leap in solving protein structures"
    - Supplemental Readings: CHAI chapter 15,
      Chalmers (2010) "The singularity: A philosophical analysis"