PHI 3692: Honors AI Ethics
Spring 2024
Professor Mark Tunick
3 credits, No prerequisites
Meets MW 3:30-4:50 in AD205
Office: HC 104
Office Hours: MW 11-2 and 2-3 in HC 104. Other times by arrangement: via zoom.
Contact: tunick@fau.edu; (561) 799-8670

Description: This course will take an interdisciplinary approach to AI Ethics, drawing on science fiction, films, philosophy, political theory, social science, and work by scholars in the health care and computer science fields who explore the implications of AI technology. No prior knowledge of AI is assumed, and we will begin by learning some basics about machine learning and LLMs so that we can better address the ethical issues raised by AI.

Rapid developments in artificial intelligence (AI) and in particular recent advances in Large Language Models (LLMs) such as Chat-GPT and image generators like Dall-E 3 have led to some mind-boggling results that seem to promise revolutionary advances in technology, such as transforming health care with personalized medical treatment, revolutionizing education, identifying new drugs, reducing crime or environmental harms through predictive analytics, and vastly improving the efficiency of many professionals including authors, lawyers, engineers, musicians, artists, designers, and business managers. At the same time, they raise numerous ethical issues that will be the focus of our course.

There are speculative philosophical concerns: will AI eventually become so intelligent it will threaten humanity's very existence? If AIs someday have the capacities of humans would they be owed the same moral consideration that we give humans? Part of the course is dedicated to addressing such philosophical issues by considering whether AI robots could ever be persons with rights, or whether LLMs can genuinely understand, think and create original works. To address such issues we consider what it is to be a person and to have intelligence.

The rest of the course focuses on more immediate ethical concerns. These include the risks LLMs pose to the environment (because training models have an enormous carbon footprint), or to public health and political stability (due to their ability to spread toxic speech or misinformation); whether AI will make most of our jobs obsolete, leading to widespread unemployment and dissatisfaction; whether ChatGPT will make the college essay obsolete; who should be morally and legally responsible if a self-driving car or military drone causes an accidental death? Is it possible to have a loving relationship with an AI? Is it moral to sell or use sex robots? Who if anyone should own the patent or copyright to creative works or inventions made by LLMs? Because AI is used to analyze and mine 'big data' we will also address ethical issues associated with big data, such as concerns of bias in the use of predictive policing, predictive analytics in health care, admissions, or hiring and evaluating employees.

This course fulfills the HC Core 'Humanities B' or Global Citizenship: Ethics requirements and counts as a Humanities Distribution Elective. There are no prerequisites.

Course Objectives: To understand AI capacities, the ethical issues raised by the use of AI, and critical approaches to thinking about these issues; to understand the distinction between

humans, persons, and various forms of non-humans including robots, and the ethical implications of these distinctions; to improve one's ability to read complex ethical arguments independently, to think critically, write clearly and with precision, and develop oral skills through participation in class discussion and delivering a short presentation.

Requirements: Class will be discussion-based. It is important for students to come to class prepared to discuss the reading scheduled for that meeting. Grading is based on quizzes/discussion boards (35%), three papers of 3-5 pages each (45% total), a short class presentation on a supplementary reading (10%), and class participation (10%). Because this is a discussion-based course, attendance is important and so the participation grade will be reduced by 1.5 points for each unexcused absence beyond 2. The default grading scale is 94-100 (A), 90<94 (A-), 87<90 (B+), 84<87 (B), 80<84 (B-), 77<80 (C+), 74<77 (C), 70<74 (C-), 67<70 (D+), 64<67 (D), 61<64 (D-), <61 (F). I may utilize a curve but only if doing so would yield a higher rather than a lower grade than what the student would earn using the default scale.

Reading: 3 books have been ordered through the bookstore or can be purchased through Amazon or other sources: Cathy O'Neil, Weapons of Math Destruction (2017, ISBN 0553418831); Isaac Asimov, I, Robot (1950, ISBN 0553294385); and Kazuo Ishiguro, Klara and the Sun (2021, ISBN-10: 0593311299). All other material is available in Canvas. Reading listed under each class is to be done prior to that class meeting.
Canvas: This course makes use of Canvas, where you will find readings, background information, quizzes and discussion boards. Be sure to check Canvas regularly for updates.

Schedule/Topics [subject to minor changes: Check Canvas for updates and 'for those interested' material, some of which will be the topics for student presentations]

I. Introduction to AI and its capabilities.
This course presupposes no knowledge of computer programming, but we begin with some introductory material on A.I. and LLMs (Large Language Models), including an introductory section from a classic text on AI programming (Russell and Norvig), a discussion of new methods for creating artificial intelligence using model building rather than pattern recognition (Tenenbaum et.al.), and students are encouraged to view some training videos on LLMs.

Jan. 8: In class discussion of Searle's 'Chinese Room' Argument; excerpt from Extant Season 1: Episode 1 (17:34-18:59 and 19:58-22:16).
Reading (prior to class): Bartneck, What is AI? (2021), ch. 2 excerpts (7 pages); Primer on Ethical Frameworks (in Canvas). Students will also start to use some LLMs to create text and images and discover their abilities and limitations on their own (links in Canvas).

Jan. 10: Introduction to AI and Programming robots
Rdg: Russell and Norvig, Artificial Intelligence, ch. 2 and ch. 3 sec. 1 (24 pages).

Jan. 15 No class (M.L.K. Day)

Jan. 17: Introduction to Machine Learning
Rdg: Mark Coeckelbagh, AI Ethics (2020), ch. 6 on Data Science and Machine Learning (13 pages); Tenenbaum et.al., 'Building Machines that learn and think like people', *Behavioral and Brain Sciences* 40 (2017), excerpts (16 pages)

Jan. 22: Introduction to Large Language Models (LLMs): do they understand what they create?
Rdg: Kyle Mahowald, Anna Ivanova, Joshua Tenenbaum et.al., 'Dissociating Language and Thought in Large Language Models', arXiv:2301.06627v2 (Nov. 4, 2023)
Recommended: Linked-in Learning videos on LLMs (links in Canvas; log in using FAU credentials)

Jan. 24: Will AI Surpass Humans and threaten our existence?
Rdg: 1)Nick Bostrom, 'Are we living in a computer simulation?', *Philosophical Quarterly* 53(211):243-55 (2003), short excerpt (2 pages); 2) David Chalmers, "The Singularity: A Philosophical Analysis," *Journal of Consciousness Studies* 17(9-10):7-48 (omit 24-26 and last section) (2010); 3) Ralph Nader's blog post

II. Can robots be 'persons'? What is it to be human and have intelligence? What moral consideration is owed humans vs animals vs robots vs plants vs clones

Jan. 29: Kazuo Ishiguro, *Klara and the Sun* (2021)

Jan. 31. A.M. Turing, "Computing Machinery and Intelligence," *Mind* 59:433-60 (1950)-omit section 5 (439-442)
Discussion Board post

Feb. 5: Mary Ann Warren, Moral Status: Obligations to Persons and other living things (1997), ch. 1, 2, 4

Feb. 7: Mary Ann Warren, Moral Status, ch. 6

Feb. 12: Film: 'Her' (126 minutes), which will be screened prior to class (to be announced); Neil McArthur, "The Case for sexbots," in Danaher and McArthur, eds. Robot Sex (2018) (15 pages); and 'This Man married a Fictional character', NYT 5-24-2022

Feb. 14: Michael Hauskeller, "Automatic Sweethearts for Transhumanists," in Danaher and McArthur, eds. Robot Sex (2018)(15 pages)

Feb. 19 Stephen Petersen, "The Ethics of Robot Servitude," *Journal of Experimental and theoretical Artificial intelligence* 19(1):43-54 (2007)
Film: Star Trek Next Generation, Season 2: Episode 9, "The Measure of Man" (start at 6:46), shown in class.
Paper 1 Due

III. Can morality be programmed into a machine?

Feb. 21: Ethical frameworks: utilitarian, deontological, virtue theory; and Asimov's approach
Rdg: Isaac Asimov, I, Robot: 'Robbie', and 'Runaround'

Feb. 26: Isaac Asimov, I, Robot: 'Reason', 'Liar', 'Evidence', 'The Inevitable Conflict' (and for possible extra credit, 'Little Lost Robot')

Feb. 28: Programming moral machines
Rdg: Wendell Wallach et.al., Moral Machines, chapters 5-6 (24 pages); Wallach, 'A Conceptual and Computational Model of Moral Decision Making', *Topics in Cognitive Science* 2:454-85 (2010)(excerpts, 4 pages)

March 4, 6: no class (spring break)

4. Topics and Applications

A. Military drones and self-driving cars
March 11: Killer drones
Rdg: 'As AI-Controlled Killer Drones Become Reality, Nations Debate Limits', NYT Nov. 21, 2023; Scholz and Galliott, 'The Case for Ethical AI in the Military', in *Oxford Handbook of Ethics in AI* (2020); and Robert Sparrow, "Killer Robots," *Journal of Applied Philosophy* 24(1) 62-77 (2007). Discussion Board Post.

March 13: Self-driving cars
Rdg: Edmond Awad, et.al., "The Moral Machine Experiment," *Nature* 563:59-65 (Nov. 2018); *Brouse v. U.S.,* 83 F. Supp. 373 (1949) (1 page); 'Tesla Recalls Autopilot Software in 2 Million Vehicles', NYT Dec. 13, 2023.

B. AI and Work
March 18: Pegah Moradi and Karen Levy, 'The Future of Work in the Age of AI: Displacement or Risk-Shifting' (2020); Derek Thompson, "A World without Work," The Atlantic Monthly July/Aug 2015, online (12 pages); 'In Reversal Because of AI, Office Jobs (white-collar) are now more at risk', NYT 8-24-2023; Derek Thompson, 'Your creativity Won't Save your job from AI', The Atlantic, Dec. 1, 2022
Paper 2 Due

C. AI, big data, politics, and the threat to human autonomy
March 20: O'Neil, Weapons of Math Destruction (2017): Intro, Chs 1, 3-5

March 25: O'Neil, Weapons of Math Destruction, Ch. 6-10, Conclusion, and Afterword

March 27: Filter bubbles and echo chambers
Rdg: Pariser, The Filter Bubble, Introduction (11 pages); Efrat Nechushtai and Seth Lewis, 'What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations', Computers in Human Behavior

90:298-307 (2019); Adam Piore, 'No Big Tech Didn't make us polarized', MIT Tech Review 121(5):18-21 (2018)

April 1: Manipulation
Rdg: Kramer et. al., 'Experimental evidence of massive-scale emotional contagion through social networks', *Proceedings of National Academy of Sciences* 111(24):8788-90 (June 17, 2014); Robert Epstein and Ronald Robertson, "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections', *Proceedings of National Academy of Sciences*, Aug. 4, 2015 online (9 pages)

April 3: Deception and its impact on Human-AI relationships
Rdg: Judith Donath, 'Ethical Issues in our relationship with artificial entities', Oxford Handbook of Ethics of AI (2020); Will Knight, "Fake America Great Again: Inside the race to catch the worryingly real fakes that can be made using A.I.," *MIT Technology Review* 121(5):37-41 (2018); Tiffany Hsu, "These influencers aren't flesh and blood, yet millions follow them," *New York Times*, June 17 2019; Neudert, "Teaching propaganda how to talk," *MIT Technology Review* 121(5):72-3 (2018); 'A.I. Muddies Israel-Hamas War in Unexpected Way', NYT 10-28-2023; Film: Black Mirror, 'Be Right Back', Season 2:1 (at Netflix).

D. Ethical issues raised by LLMs

April 8: LLM Bias and risks of harm
Rdg: Dan Mcquillan, 'Chat GPT is a bullshit generator', Vice, Feb. 9, 2023; Eily Bender, Timnit Gebru et.al., 'On the Dangers of Stochastic Parrots: Can LMs be too big?', FAccT '21, March 3-10, 2021; Luigi De Angelis et.al., 'Chat GPT and LLMs: New AI driven infodemic threat in public health', Frontiers Public Health 25 (April 2023); Open AI, 'GPT-4 System Card', March 23, 2023.

April 10: LLMs and Education
Rdg: Perkins, 'Academic Integrity Considerations of AI LLMs in a Post-Pandemic World', Journal of University Teaching and Learning Practice 20(2) (2023); Meyer et.al., 'Chat GPT in academia: opportunities and challenges', Biodata Mining (editorial), July 13, 2023; Stephen Marche, 'The College Essay is Dead', The Atlantic, Dec. 6, 2022; Anna Strasser, 'On pitfalls [and advantages] of sophisticated LLMs', ArXiv.org 2303.17511.pdf (2023)
Discussion Board Post.

April 15: LLMs and copyright
Rdg: Thaler v. Perlmutter, __ F. Supp. 3d. __ (Dist. Ct, D.C., 2023); Christopher Zirpoli, Congressional Research Service, 'Generative AI and Copyright Law', Sept. 29, 2023; Appel et.al., 'Generative AI has an intellectual property problem', Harvard Business Review, April 7, 2023.

E. AI and Healthcare
April 17: Rdg: Eric Topol, Deep Medicine: How AI can Make Healthcare Human Again (2019), ch. 1-3 (58 pages) ; Pam Belluck, 'Nuns offer Clues to Alzheimer's and Aging', *New York Times*, May 7, 2001; Doctors wrestle with AI in patient care, citing lax oversight, New York Times,

10/30/2023.

April 22: Rdg: Topol, Deep Medicine, chs. 11-13 (77 pages); Daisuke Wakabayashi, "Google and University of Chicago Are Sued Over Health-Data Sharing," *New York Times* June 26, 2019; David Lazer et.al., 'The Parable of Google Flu: Traps in Big Data Analysis', *Science* 343:1203-5 (13 March 2014);
Paper 3 due

Additional notes:

Attendance Policy: Students are expected to attend all of their scheduled University classes and to satisfy all academic objectives as outlined by the instructor. The effect of absences upon grades is determined by the instructor, and the University reserves the right to deal at any time with individual cases of non-attendance. Students are responsible for arranging to make up work missed because of legitimate class absence, such as illness, family emergencies, military obligation, court-imposed legal obligations or participation in University-approved activities. Examples of University-approved reasons for absences include participating on an athletic or scholastic team, musical and theatrical performances and debate activities. It is the student's responsibility to give the instructor notice prior to any anticipated absences and within a reasonable amount of time after an unanticipated absence, ordinarily by the next scheduled class meeting. Instructors must allow each student who is absent for a University-approved reason the opportunity to make up work missed without any reduction in the student's final course grade as a direct result of such absence.

Policy on Accommodations In compliance with the Americans with Disabilities Act Amendments Act (ADAAA), students who require reasonable accommodations due to a disability to properly execute coursework must register with Student Accessibility Services (SAS) and follow all SAS procedures. SAS has offices across three of FAU's campuses -- Boca Raton, SU 131 (561-297-3880); in Davie, LA 131 (954-236-1222); in Jupiter and all Northern Campuses, SR 111F (561-799-8585). Disability services are available for students on all campuses. For more information, please visit SAS website at www.fau.edu/sas/.

Counseling and Psychological Services (CAPS) Center Life as a university student can be challenging physically, mentally and emotionally. Students who find stress negatively affecting their ability to achieve academic or personal goals may wish to consider utilizing FAU's Counseling and Psychological Services (CAPS) Center. CAPS provides FAU students a range of services – individual counseling, support meetings, and psychiatric services, to name a few – offered to help improve and maintain emotional well-being. For more information, go to http://www.fau,edu/counseling/

Academic Integrity Policy: Students at Florida Atlantic University are expected to maintain the highest ethical standards. Academic dishonesty is considered a serious breach of these ethical standards, because it interferes with the university mission to provide a high quality education

in which no student enjoys an unfair advantage over any other. Academic dishonesty is also destructive of the university community, which is grounded in a system of mutual trust and places high value on personal integrity and individual responsibility. Harsh penalties are associated with academic dishonesty. For more information, see University Regulation 4.001 and http://www.fau.edu/divdept/honcol/students/honorcode.htmlLinks to an external site.
Policy on use of Artificial Intelligence (AI): Students are not permitted to use AI (such as ChatGPT, PaLM2, Grammarly-Go, any other LLMs, etc.) in working on a graded assignment for a class (e.g., written work such as papers, quizzes, discussion board posts, or any other assignment), unless explicitly permitted to do so by the instructor. See the Honors College Policy on the Use of AI in Courses.

Classroom Etiquette Policy: To enhance and maintain a productive atmosphere for education, personal communication devices, such as cellular telephones and pagers, are to be disabled in class sessions.
In this class we address many controversial social issues. The goal is for students to form their own judgements and develop their critical thinking skills. Students are to be respectful of other students and recognize that reasonable people can take different positions on the issues we address.
Policy on Recording in Class: by state law, audio or video recordings of class lectures is permitted only for personal educational use and may NOT be published. Publication, which refers to circulating, sharing, or distributing with anyone (including classmates) or on social media or other media formats is by law subject to penalties up to $200,000. In addition, failure to adhere to this policy may constitute a violation of the honor code. Recording of class discussions is not permitted unless the student has an accommodation granted by Student Accessibility Services. Students who request recording of class lectures or discussions under the Americans with Disabilities Act must contact Student Accessibility Services to obtain such permission or accommodation, and must otherwise comply with the requirements of SAS. Information for the SAS is available at http://www.fau.edu/sas/.