| Course Code | Course Title |
| --- | --- |
| **COMP 1944** | **Artificial Intelligence Ethics** |

Course Description

AI is disrupting every sphere of our work and lives, bringing unprecedented risks to society. This introductory course surveys the explosive area of artificial intelligence ethics, illuminating relevant AI concepts with no prior background needed. Fake news bots. AI driven social media displacing traditional journalism. Drone warfare. Elimination of traditional jobs. Privacy-violating advertising. Monopolistic network effects. Biased AI decision/recognition algorithms. Deepfakes. Autonomous vehicles. Automated hedge fund trading. No area remains untouched. Policy think tanks, governments, and tech companies around the world have started paying serious attention to AI ethics. How will human civilization survive the rise of AI? What are the new rules? What are the ethical frameworks needed to avoid extinction? What are engineers' and entrepreneurs' ethical responsibilities? Previous Course Code(s): COMP 4901M

List of Topics

**Overview**. Fairness, accountability, and transparency in society, AI and machine learning, the impact of AI and automation upon labor and the job market

**Information disorder.** Misinformation vs disinformation vs malinformation, deepfakes, chatbots, and how they disrupt society

**Algorithmic bias, unconscious bias, and inductive bias.** Relationships between three different foundational kinds of bias and the serious social consequences

**AI ethics methodologies.** Survey of approaches to formulating AI ethics methodologies

**Weak AI, strong AI, and superintelligence.** Contrasts between different senses and levels of "AI". that impact human-machine interaction and society in very different ways

**Conscious AI.** Artificial mindfulness and its societal impact

**Prescriptive vs descriptive AI ethics, and deontological vs consequentialist vs virtue AI ethics.** Relates classic ethics philosophy to the problem of AI ethics, and discusses why purely rule-based AI ethics will fail

**Artificial moral cognition.** Embedding ethics into AIs themselves, and the illusion of explainability

**Privacy, safety, security.** Surveillance capitalism, identity theft, artificial gossips

**AI-driven framing and narratives.** The role of social media, recommendation engines, and search engines in computational propaganda and artificial storytellers

**Constructive and creative AI.** Are machines more creative than humans? Computational creativity and its impact on society and culture

**Empathetic AI.** Affective computing and artificial intimacy

**AGI safety.** Is the future of humans extinction, to be pets in a zoo, to upload, or to merge with AIs?

Textbooks

**Artificial Children**, 2022, Dekai Wu

Reference books

**IEEE Ethically Aligned Design**

**Artificial Intelligence: A Modern Approach, 4th ed.**, 2020, by Stuart Russell and Peter Norvig

**The Fundamentals of Ethics, 5th ed**, 2020, by Russ Shafer-Landau

Grading Scheme

20% in-class questionnaires/quizzes
10% midterm exam
20% written assignments
25% group presentation
25% course participation

Course Intended Learning Outcomes

Upon completion of this course, students are expected to be able to do the following:

1. Compare different frameworks for AI ethics, including IEEE Ethically Aligned Design (ST2), and be familiar with the many types of misuse of AI technology (SA2)

2. In new scenarios where AI impacts society, evaluate the limits of deontological rule-based AI ethics; analyze intended and unintended consequences in line with consequentialist AI ethics, social well-being metrics, and AI for social good; and analyze the societal role of virtue AI ethics (ST2, SA1)

3. Propose designs for embedding values into autonomous systems and artificial moral cognition, particularly with regard to the ethics of emotional AI, empathetic AI, and affective computing (ST1, ST2, SA2)

4. In new scenarios where AI impacts society, analyze the AI and machine learning's fairness, responsibility and accountability, and evaluate what degree of transparency and explainability is possible (ST1, ST2, SA1, SA2)

5. Recognize weaponization of information and exploitation of unconscious biases, and propose training dataset design policies to avoid algorithmic bias and discriminatory outcomes (ST1, ST2, SA1, SA2)

6. Analyze the social consequences of alternative approaches of personal data rights and individual access control, and recognize the risks of surveillance capitalism (ST1, ST2, SA2)

7. Explain the risks of autonomous weapons, analyze the tradeoffs, and contrast policy proposals (ST1, ST2, SA2)

8. Analyze AI safety options in the coming eras of strong AI and artificial superintelligence (ST1, ST2, SA2)

Assessment Rubrics

N/A