Spring 2021
Case Western Reserve University
Department of Philosophy

TuTh 1:00 PM – 2:15 PM
Tinkham Veale University Center 280F
(Inamori International Center)
Zoom meeting location (first 2 weeks):
See Canvas course page

Dr. Daniel Rosiak
Postdoctoral Research Associate
Inamori International Center for Ethics
& Excellence

dhr30@case.edu
Office hours: TuTh 2:30 PM – 4PM
(by appointment)

AI Ethics 2022

Course Description

Artificial Intelligence (AI) is a hot topic these days, not least on account of the many powerful and subtle impacts it is having on so many aspects of our lives, and will likely increasingly have in the future. AI can already save and improve lives by improving medical diagnosis, making new medical and scientific discoveries, better predicting extreme weather events and anomalies, protecting and monitoring natural environments, preventing suicide, conducting fact checking, improving food production and crop management – and so much more. And this is just to speak of the present. AI can save and improve lives, but it also poses profound dangers, for instance with respect to autonomous weapons systems, surveillance and control, behavioral nudging, attention engineering, erosion of truth (e.g., via deepfakes), safety-critical applications, perpetuating societal biases, impacts on employment – and so much more. Again, this is just to speak of the present. AI's diverse applications (both present and future) raise a host of concerns of the highest order. For these and other concerns having to do with natural extensions of such capabilities as we look to the future, AI presents us with a myriad of ethical challenges. The object of this course is to address some of these challenges.

While the scope of impact of such technologies is impressive, there is also a great deal of hype and sweeping generalizations about AI. Precisely on account of the gravity of the concerns associated with AI technology, we would do well to avoid such sweeping claims and ground our investigations in an understanding of the actual state of AI technologies and their possible extensions in the future. This course will accordingly tether its exploration of ethical matters in a nuanced appreciation of, and familiarity with, the core existing capacities of AI. The student will learn – without needing prior knowledge of programming or computer languages – the basics of the major techniques and capacities that comprise AI, grounding discussion of ethical matters in real-world applications. Building on this knowledge, we will then extrapolate future capacities and consider their associated ethical concerns.

Readings

All required course readings will be supplied electronically via Canvas; if you like to have a physical copy, you can purchase any of these texts, but it is not required that you do so.

Selections from

- *Artificial Intelligence: A Modern Approach* (4th edition) (Russell and Norvig)
- "Rosiak Notes on Ethics & AI" (Rosiak)
- *Fairness and Machine Learning* (Solon Barocas, Moritz Hardt, Arvind Narayanan)
- *Weapons of Math Destruction* (Cathy O'Neil)
- *The Alignment Problem: Machine Learning and Human Values* (Brian Christian)
- *Algorithms of Oppression* (Safiya Noble)
- *Superintelligence* (Nick Bostrom)
- *AI Superpowers* (Kai-fu Lee)

Class During a Pandemic

As of January 08, 2022, the university has said that we will be conducting class remotely for at least the first two weeks. As long as we are remote, we will be using Zoom for class meetings. For these meetings, students are strongly encouraged to have their cameras on as much as possible, as this better facilitates engagement and class discussion. Please log into the Zoom meetings by class start time, not at or after: this course has lots of interesting material to cover, so every minute counts!

Graded Assignments

1. **Weekly Problem-sets**: Each week (except for the week of the midterm, the week of the final, and the weeks of your final presentations) students will have a problem-set dedicated to further exploring the material covered in the class meetings, readings, and further course material of that week. Altogether, these problem-sets make up 25% of your final grade. There are 11 of these total, so each is worth ~2% of your grade in the course. Other than the designated exception weeks, these are due each Friday of the week, at 11:59 PM.

2. **Midterm Paper**: Each undergraduate student* will write a 1250-1750 word midterm paper engaging with one of the topics covered in the first half of the course. Works Cited and title pages do not count towards the required word count. Details of what is required for this assignment can be found on Canvas. The paper must be received by 11:59pm on February 20.

*NOTE: Any graduate or professional students taking this course will not write this midterm paper, but instead will be required to write a single, longer paper that delves more deeply into the course material, in place of the two shorter papers required of the undergraduates. The submission of an interim draft will also be required during the same time the midterm is due for the rest of the class.

3. **Final Paper**: Each student will write an end-of-term paper (2000-3000 words for undergraduates, 4000-5000 words for graduate and professional students).
Paper topics will be chosen by students and approved by the instructor by 11:59pm on April 8. The paper must be received by 11:59pm on April 29.

4. **Final Presentation**: Each student will give a final presentation relating to the topic explored in his or her end-of-term paper. These presentations will take place during the last two weeks of the semester (a sign-up sheet with presentation dates will be provided closer to this time). Students should endeavor to devise a memorable way to grab and retain their classmates' attention and challenge them to engage the material on a deeper and more personal level. Creativity is encouraged for these presentations. These final presentations provide the opportunity for students to get feedback from their classmates and instructors on the topic covered in their final paper.

5. **Participation**: Students' course grades will also be affected by their overall level and quality of class participation. Students must do the reading before class, be active listeners, and contribute to the class discussion. A discussion forum has been set up on Canvas where students are encouraged to post questions or comments about the material and continue discussions outside of class. Engagement on this discussion forum is an additional factor of class participation.

Grades

25% = Problem-sets (11 total)
25% = Mid-term paper
15% = Final presentation
25% = Final paper
10% = Class participation (quality and quantity)

Course Schedule

*****Note that the reading material listed under a particular class meeting will be discussed ***that day*** – i.e., you should have read the material ***prior to*** the start of that class meeting. All of the reading and course assignment materials can be found on Canvas.*

Week 1 Introduction to the Course and Overview of AI Ethics Issues

- What is AI? Narrow AI / AI in present vs. Strong AI / AGI / AI in the future
- Overview of Dangers / Risks (using the 2 Concept Maps)
- Glimpse into some of the Potential / Opportunities
- Overview of Course

Readings

Jan 13

- Introduction to *Artificial Intelligence: A Modern Approach* (Chapter 1)
- Review the 2 Concept Maps (Narrow AI & Strong AI)

Assignment

Due Jan 14 at 11:59 PM

- Problem-Set #1

1st half: The Scope of Narrow AI / AI in the present

Overview of 1st half: Weeks 2-5: Zeroing in on what AI is: what problem is AI the answer to? what can AI do? what can't it do? A tour, via real-life examples and/or simple illustrative cases, unpacking the 4 main areas of AI, and grounding the introduction of some of the main issues in terms of a knowledge of how these arise in real applications; Week 6: Broad overview of issues (potentially of ethical import) internal to the AI tools

Week 2 Background Fundamentals on Intelligent Agents; The First Area (Problem-solving)

1. Problem-solving
   a. search problems & optimization
   b. constraint satisfaction problems (+ types of constraints)
   c. limitations

Readings

Jan 18

- *Artificial Intelligence: A Modern Approach* (Chapter 2)

Jan 20

- Rosiak Notes (Parts 1 and 2)

Assignment

Due Jan 21 at 11:59 PM

- Problem-set #2: exploring features and identifying limitations in two examples

Week 3 The Second Area (Reasoning/Planning)

2. Reasoning/Planning
   a. classical
   basics of goal-directed reasoning, formal languages, and intelligent agents ; combining best of formal and natural languages

   b. probabilistic (under uncertainty)
   overview of basics
      -basics on types of uncertainty and how to deal; intro to Bayesian networks
      -simple example with diagnosing cavity
      -complicated example with evaluating car insurance application and determining premium
      -dynamic example: finding the most likely sequence of events given some state
   combining beliefs and decisions
      -basics of utility theory
      -determining risk (returning to insurance policy example)

-public policy airport example (weighing costs against benefits in picking new site for an airport)
-ice-cream example: decision making when there is uncertainty about the utility function whose expected value is to be optimized
-off-switch game example involving determining when a machine helper should defer to the human
-glimpse into complex decision-making: bandit problems

<u>multiagent planning</u>

-basics of game theory through a coordination problem
-assistance games: paperclip game example
-resource allocation problem

c. limitations

Readings

Jan 25

- Rosiak Notes (Part 3)

Jan 27

- Rosiak Notes (Part 4)

Assignment

Due Jan 28 at 11:59 PM

Problem-set #3

Week 4 The Third Area (Machine Learning)

First, covering leftover material from Week 3 – on decision theory and decision networks...then resuming with:

3. Learning from Examples (Machine Learning)

<u>overview</u>

-The basic theory of ML, and typical steps

<u>3 main types of learning (based on the 3 types of feedback accompanying inputs)</u>

a. supervised learning
-example of image classification
-example of suicide prediction

b. unsupervised learning
-example of anomaly detection; sentiment analysis

c. reinforcement learning
-example of traffic control & self-driving; personalized recommendations
-exploration
-reversing the process: Apprenticeship learning, or learning how to behave well given observations of expert behavior (robotics example)
-issue: catastrophic forgetting (use autonomous vehicle example)

Readings

Feb 1

- Rosiak Notes (Part 5)

Feb 3

- Rosiak Notes (Part 6)

Assignment

Due February 4 at 11:59 PM

Problem-set #4: exploring features and identifying limitations & broader dangers in two examples

---

Week 5 Continuing with ML; plus, the Fourth Area (Communicating/Perceiving/Acting)

First finishing up with ML:

deep learning more generally

-briefly describe a few neural network examples: facial recognition & weather forecasting

-mining data from social media platform

-GANS: deep fakes

issues

-trust, interpretability/explainability

-problem of long tail of user inputs

The Fourth Area:

4. Communicating/Perceiving/Acting

3 important areas

a. NLP

-information extraction examples

-returning to sentiment analysis example

-urgency detection

-some state-of-the-art applications

b. vision

-3D rendering the metaverse

-classifying images with convolutional neural networks (Tesla/Mobileye sensing example)

-classifying what people are doing by analyzing video

-return to deepfakes

c. robotics

-types of robots

-what sorts of problems is robotics solving?

-example with self-supervised robot (i.e., robot that collects its own training data, with labels)

-military application

-safe exploration

Readings

Feb 8

- Rosiak Notes (Part 7)
- Selections from *Artificial Intelligence: A Modern Approach* (On ML)

Feb 10

- Rosiak Notes (Part 8)

Assignment

Due February 11 at 11:59 PM

Problem-set #5: exploring features and identifying limitations & broader dangers in two examples: example of NLP-based sentiment analysis and its uses; deepfakes with GANS

Week 6 Finishing up with the Fourth Area; plus, Broad Overview of Internal Issues re Existing Techniques

Fourth Area continued from last week:

more advanced considerations

-overview of coordination problem (once robots are more integrated in human environments): game of predicting future behavior via autonomous vehicle

-indirectly learning what humans want by learning desired cost functions they should optimize from inputs

-imitation learning: imitating effective human behavior

-adversarial training

-general considerations with human-machine interfaces: example with brain-machine interface controlling robot arm

issues

-with imitation learning: generalization to new states

Broad Overview of Internal Issues:

Risks (via examples)

- opacity
- bias (types and trade-offs)
- goal blindness
- rigidity
- edge cases
- bad memory
- difficulty generalizing / brittleness
- quantifying uncertainty

Opportunities

- improvements in smooth functioning of existing systems
- solving hard problems, diagnosing, etc.
- potential benefits of automation, smart decision-making,
- identifying & rectifying instances of bias / injustice / holes in knowledge base

Readings

Feb 15

- Rosiak Notes (Part 9)

Feb 17

- Rosiak Notes (Part 10)

Assignment

Midterm Due Feb 20 at 11:59 PM

2nd half: Dangers and Opportunities in AGI; or how it may be in the future

Overview of 2nd half: Weeks 7-9: a deeper dive into the ethical challenges presented by today' s AI ; weeks 10-12: exploring the ethical challenges presented by future trajectories; weeks 13-14: student presentations

---

Week 7 Humans weaponizing against other humans

Risks / Issues

- Data privacy erosion & attacks
- Behavioral nudging / Attention Engineering
- Surveillance & control: social manipulation / predictive administration dystopia
- Military uses: lethal autonomous weapons
- Escalating geopolitical conflicts / international tensions

Opportunities

- New economy of self-sovereign data creators; new modes of preserving privacy and de-identification
- Beneficial nudging / benefits of personalization
- Reduction of bureaucratic elements
- More "humane" conflicts

Readings

Feb 22

- Selections from *AI Superpowers* (Ka-fu Lee)

Feb 24

- Selections from *I, Robot*
- Selections on self-sovereignty (tbd)

Assignment

Due February 25 at 11:59 PM

Problem-set #6: Case study

Week 8 Intensification of existing social problems

Risks / Issues

- Fairness issues (different fairness frameworks; COMPAS recidivism scoring example) revisited & a closer look at bias: social injustice acceleration via learned historical bias & predictive policing
- Trust and transparency: responsibility vacuums & explainable AI

Opportunities

- improved transparency, explicitness, and accountability

Readings

March 1

- Selections from *Fairness & Machine Learning*

March 3

- Selections from *Weapons of Math Destruction*

Assignment

Due March 4 at 11:59 PM

Problem-set #6: Case study

Spring Break March 7-11

Week 9 Introduction to Superintelligence and its issues

Readings

March 15

- Selections from Nick Bostrom's *Superintelligence*

March 17

- Selections from *Superintelligence*

Assignment

Due March 18 at 11:59 PM

Problem-set #7: Case Study

Strong AI / AGI dangers & opportunities / AI in the future

Week 10 A Deeper Look at Superintelligence & Value-loading

Broad overview of value-loading problem & related problems

- Envelope problems (e.g., what values to include)
- Direct vs. Indirect approaches
- Direct specification difficulties
- Unpredictability of indirect normativity
- Unanticipated ways of fulfilling final goal

Value alignment & misalignment

Risks

- soft treacherous turn
- value divergence

Opportunities

- quality of life enhancement
- more just social outcomes

Extreme Possibilities re humanity

Risks

- modes of perverse instantiation:
  - wireheading
  - “that’s not what we meant”
  - inscrutability of final goals
  - out-parented (final goal “re-alignment”)
- hard treacherous turn (strategic vs. unanticipated)

Opportunities

- “virtue assistants”

Extreme Possibilities re AGI: Robot Rights issues

Risks

- stunting
- value divergence
- exploitation / enslavement
- mind crime

Opportunities

- improved employment opportunities / quality of employment
- unparalleled insights from simulations

Readings

March 22

- Selections from *Superintelligence*

March 24

- Selections from *Superintelligence*
- Selections from *The Alignment Problem: Machine Learning and Human Values*

Assignment

Due March 25 at 11:59 PM

Problem-set #8: Answer questions about value-loading

Week 11

Existential-level threats / changes

Existential-level changes

Risks

- runaway optimization
- extreme competition
- malevolent AGI
- destructive species dynamics

Opportunities

- human-AI symbioses

Lifestyle change dangers & opportunities

Risks

- future of capital / economic challenges
- changes to body / human augmentation

Opportunities

- future of capital / economic challenges
- changes to body / human augmentation

Readings

March 29

- “Future of Humanity” (Bostrom); article on bioengineering (tbd)

March 31

- Article on Future of Capital (tbd)

Assignment

Due April 1 at 11:59 PM

Problem-set #11: Answer questions about future of capital/lifestyle changes

---

Week 12

Summing up current issues facing us...

Weeks 13-14: Student presentations

Final Paper Due April 29 at 11:59 PM