UNIVERSITY of GUELPH

PHIL 6400
Data Science Ethics
Winter 2023

****Please check Syllabus and Announcements on CourseLink for the most up-to-date information****

Schedule: Tuesdays, 2:30 – 5:20 PM in MCKN 119
Professor: Joshua August Skorburg (I go by "Gus")
E-mail: skorburg@uoguelph.ca
Office Hours: By appointment on Teams on in MCKN 336

Course Description

This course will explore the broad philosophical implications (ethical, legal, social, political, epistemological, etc.) of recent developments in data science, artificial intelligence, and machine learning. We will consider important themes from the philosophy of science, the philosophy of technology, ethics, and social/political philosophy as these relate to recent developments data science, artificial intelligence, and machine learning.

Required Resources

Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence.* Yale University Press. ISBN: 9780300264630. All other required readings will be posted as .pdfs on Courselink.

Learning Outcomes

By the end of this course, you should be able to:

1. Identify and critically evaluate a wide range of philosophical issues related to data science, artificial intelligence, and machine learning.

2. Produce professional, academic writing and commentary related to philosophical implications of data science, artificial intelligence, and machine learning.

3. Contribute to ongoing academic and popular discussions about the ethics of emerging technology.

Assessments

You need to develop the skills that are most important to launching and sustaining your career: scholarly research and the ability to present it. To that end, you will be assessed on course engagement, two short papers (2,000 words each), and one presentation, as follows:

1. Presentation (20% of final grade)
Over the course of the term, each of you will make an in-class presentation of around 20 minutes. These presentations are low-key and *simply meant to help stimulate discussion on the topics at hand*. The content is up to you, but in general, the presentations should identify and elaborate upon a central theme, argument, concept, etc. from the week's readings. It may be helpful to think of your presentation as a cross between a teaching demonstration and a conference presentation. A few days before your scheduled presentation date, please send a brief e-mail to Gus with a few sentences about your plan for the presentation so that we don't have too much overlapping content.

Taking the first week's reading as an example, a good strategy for a presentation could be any of the following:

- Briefly summarizing the two main ways that Winner thinks values, principles, and power-relations are embedded within socio-technical systems and then developing an argument for where deepfake technology ought to fit in this framework, or developing an argument for why deepfakes don't easily fit within Winner's framework.
- Reviewing newer literature on deepfakes and developing an argument about how more recent findings might support or undermine Rini's claims about the epistemic threats of deepfakes, perhaps providing specific examples of how deepfakes are currently being used.
- Summarize relevant technical literature on Generative Adversarial Networks to develop an argument that ethicists have misunderstood/oversimplified/overstated/overlooked/underappreciated, etc., some important aspect of the technology and its applications.
- And in general, you can always summarize the readings from the "additional literature" and connect them with themes from the required readings. In any case, you will ideally use your presentations as an opportunity to bring your unique research interests and background to bear on the assigned readings.

2. Research Paper (50% of final grade)
The central assessment for this course is a research paper of around 3,000 words. Arguably, editing and revising your writing is the most important (and most difficult!) part of scholarly research. So, the research paper will be assessed progressively throughout the course. You will first prepare a thesis statement, of a few sentences, due around Week 8. Next, you will submit an abstract or extended outline (~250 words). Around Week 10 or 11, you will submit your first draft. Then, in the style of a peer-reviewed journal submission, I will provide a "referee report" on the draft and you will then "revise and resubmit" your final draft at the end of the term. Ideally, your research paper could serve as the basis for a conference submission, thesis/dissertation chapter, etc. The tentative due date is Tuesday, April 11.

3. Engagement (30% of final grade)
This includes showing up on time, speaking up in class, paying attention to what the other students have to say, taking notes, asking questions, etc. **Engagement also includes weekly responses, due Sunday evening about the readings for the upcoming week.** These responses should be

around 250 words. The content of the response is up to you, but all responses should deal directly with the reading assigned for that week. You may want to choose a sentence or paragraph you found especially provocative, difficult, or remarkable, then explain why you found it provocative, difficult, or remarkable. Alternatively, you may want to argue that one of the authors is right or wrong in making some particular claim. Or you may wish to connect two passages in the reading that illuminate each other. You may even just pick a passage and ask questions about it. Your responses will guide what we address in class, and I will sometimes quote from them, so please take them seriously.

*always check CourseLink for most up-to-date information on scheduling, readings, assignments, etc.

**Please note that the "additional literature" for each week is meant to serve two purposes. First, it provides materials for the presenters to draw from in the presentations. Second, it provides jumping off points for students who want to write about the week's topics in their research papers. Otherwise, they are totally optional, and no component of your grade directly depends on having read them.

UNIT I: Philosophy of Technology

Week 1: January 10
Technological Neutrality

Assignments: None
Presentations: None
Readings

- Winner, L. (1980). Do artifacts have politics?. *Daedalus*, 121-136.
- Rini, R. (2020). Deepfakes and the epistemic backstop. *Philosopher's Imprint*, 1-16.

Week 2: January 17
Technological Neutrality (cont'd)

Assignment: Weekly Response 1 by 11:59 PM on Sunday January 15
Presentations:

Readings

- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2)
- Scott, J.C. (1998). *Seeing Like A State,* Chapter 1. (42 pgs)

Additional Literature

- Floridi, L., & Taddeo, M. (2016). What is data ethics?. *Philosophical Transactions of the Royal Society of London Series A*, *374*: 1-5.

- Moor, J. H. (2005). Why we need better ethics for emerging technologies. *Ethics and Information Technology, 7*(3), 111-119.

Week 3: January 24

Addiction by design

Assignment: Weekly Response 2 by 11:59 PM on Sunday, January 22

Presentations:

Reading

- Introduction, Chs. 1-3 from Natasha Dow Schüll's (2012) *Addiction by design*
- Alfano, M., Carter, J. A., & Cheong, M. (2018). Technological seduction and self-radicalization. *Journal of the American Philosophical Association*, *4*(3), 298-322.

Additional Literature

- Aagaard, J. (2021). Beyond the rhetoric of tech addiction: why we should be discussing tech habits instead (and how). *Phenomenology and the Cognitive Sciences, 20*(3), 559-572.
- Selections from Vallor (2016): New social media and the virtue of self-control
- "How casinos enable gambling addicts" (John Rosengren, December 2016, *The Atlantic*)
- "Addicted to your iPhone? You're not alone" (Bianca Bosker, November 2016, *The Atlantic*)

UNIT II: Ethics of AI

Week 4: January 31

Privacy & Human Rights

Assignment: Weekly Response 3 by 11:59 PM on Sunday, January 29

Presentations:

Reading

- Nissenbaum, H. (2011). A contextual approach to privacy online. *Daedalus*, *140*(4), 32-48.
- Amnesty International (2019) *Surveillance Giants: How the business model of Google and Facebook threatens human rights* (51 pages)

Additional Literature

- Rachels, J. (1975). Why privacy is important. *Philosophy & Public Affairs 4*(4): 323-333
- Selinger, E., & Hartzog, W. (2014). Obscurity and privacy. In R*outledge Companion to Philosophy of Technology* (Joseph Pitt & Ashley Shew, eds) 20 pgs.
- Bogen, M., & Rieke, A. (2018). Help wanted: an examination of hiring algorithms, equity. and bias. *Upturn*. (47 pages).
- Metcalf, J., & Moss, E. (2019). Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research: An International Quarterly*, *86*(2), 449-476.

<u>Week 5: February 7</u>
Corporate Research

Assignment: Weekly Response 4 by 11:59 PM on Sunday, February 5
Presentations:

Reading

- Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788-8790
- Meyer, M. N. (2015). Two cheers for corporate experimentation: The A/B illusion and the virtues of data-driven innovation. *Colo. Tech. Law Journal*, *13*, 274-327

Additional Literature

- Bird, S., Barocas, S., Crawford, K., Diaz, F., & Wallach, H. (2016). Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning* (4 pgs).
- Grimmelmann, J. (2015). The law and ethics of experiments on social media users. *Colo. Tech. Law Journal*, *13*, 219.
- "OK Cupid plays with love in user experiments" (Molly Wood, 2014, *New York Times*)
- Boyd, D. (2016). Untangling research and practice: What Facebook's "emotional contagion" study teaches us. *Research Ethics, 12*(1), 4-13.
- Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society, 3*(1), 2053951716650211.

<u>Week 6: February 14</u>
Bias and fairness

Assignment: Weekly Response 5 by 11:59 PM on Sunday, February 12
Presentations:

Reading

- Fazelpour, S., Lipton, Z. C., & Danks, D. (2022). Algorithmic fairness and the situated dynamics of justice. *Canadian Journal of Philosophy, 52*(1), 44-60.
- Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass, 16*(8), e12760

Additional Literature

- Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*, 141-163.
- "Bias" Chapter 3 in Zerilli, J. (2021). *A citizen's guide to artificial intelligence*. MIT Press.

Week 7: February 21
WINTER BREAK NO CLASS

Week 8: February 28
Transparency & Explanation

Assignment: Weekly Response 6 by 11:59 PM on Sunday, February 26; THESIS STATEMENT DUE Monday February 27 by 11:59 PM
Presentations:

Reading

- Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3(1),* 2053951715622512
- Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv:1606.03490.*

Additional Literature

- Zednik, C. (2019). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, 1-24.
- Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev.*, 87, 1085.
- Goodman & Flaxman (2016). Algorithmic decision-making and a "right to explanation" *arXiv* (9 pgs)

Week 9: March 7
Responsibility & Accountability

Assignment: Weekly Response 7 by 11:59 PM on Sunday, March 5; EXTENDED ABSTRACT/OUTLINE DUE Wednesday March 8 by 11:59 PM
Presentations:
Reading:

- Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics & Information Technology, 6*(3): 175-183.
- Rubel, A., Castro, C., & Pham, A. (2019). Agency Laundering and Information Technologies. *Ethical Theory and Moral Practice*, *22*(4), 1017-1041.

Additional Literature:

- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33-44).

- Leonelli (2016) Locating ethics in data science: Responsibility and accountability. *Philosophical Transactions of the Royal Society of London Series A*, *374*: 1-12.
- Gunkel, D. J. (2017). Mind the gap: responsible robotics and the problem of responsibility. *Ethics and Information Technology*, 1-14.
- Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, *18*(4), 299-309.
- Sparrow, R. (2007). Killer robots. *Journal of applied philosophy*, *24*(1), 62-77.

UNIT III: Politics of AI

Week 10: March 14
LLMs and the alignment problem

Assignment: Weekly Response 8 by 11:59 PM on Sunday, March 12
Presentations:
Reading:

- Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and machines, 30*(3), 411-437.
- Kasirzadeh, A., & Gabriel, I. (2022). In conversation with Artificial Intelligence: aligning language models with human values. *arXiv preprint*: 2209.00731.

Additional Literature:

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).
- Philosophers on GPT-3: https://dailynous.com/2020/07/30/philosophers-gpt-3/
- Webb, T., Holyoak, K. J., & Lu, H. (2022). Emergent Analogical Reasoning in Large Language Models. *arXiv preprint arXiv*:2212.09196.

Week 11: March 21
Atlas of AI 1

Assignment: Weekly Response 9 by 11:59 PM on Sunday, March 19; FIRST DRAFT DUE Wednesday March 22 by 11:59 PM
Presentations:
Reading:

- *Atlas of AI:* Intro, Chs. 1-2

Additional Literature:

- McKibben, B. (2022) - Could Google's Carbon Emissions Have Effectively Doubled Overnight? *The New Yorker.*

Week 12: March 28
Atlas of AI 2

Assignment: Weekly Response 10 by 11:59 PM on Sunday March 26
Presentations:
Reading:

- *Atlas of AI:* Chs. 3-4

Additional Literature:

- Fourcade, M., & Healy, K. (2017). Seeing like a market. *Socio-economic review, 15*(1), 9-29.

Week 13: April 4
Atlas of AI 3

Assignment: Weekly Response 11 by 11:59 PM on Sunday April 2
Presentations:
Reading:

- *Atlas of AI:* Chs. 5-6, Conclusion, Coda

Additional literature:

- Birhane, A., Isaac, W., Prabhakaran, V., Díaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the People? Opportunities and Challenges for Participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1-8.
- Gabriel, I. (2022). Toward a Theory of Justice for Artificial Intelligence. *Daedalus, 151*(2), 218-231.

FINAL DRAFT DUE DATE: Tuesday, April 11