PHL 377H1F: Ethical Issues in Big Data
Fall 2024, University of Toronto, St. George

Instructor:	Matthew Delhey	Time:	Tue 3–6
Office hours:	Tue 2–3 and by appointment	Location:	SS 1074
Email:	matt.delhey@mail.utoronto.ca		

1 Course Description

In this course, we will survey a rapidly growing field of philosophy: the *ethics* of AI, big data, and other digital and algorithmic technology. Our focus will lie in exploring the three main areas that have developed within this field: (1) fairness, (2) privacy, and (3) explainability. Because research in the ethics of AI is nascent and interdisciplinary, our readings will draw from recent journal articles written by a diverse body of scholars, including philosophers, computer scientists, legal scholars, statisticians, media theorists, and social scientists. Nevertheless, we will pay special attention to the contributions of philosophers to this field, with the goal of strengthening critical thinking concerning the increasing application of AI to social life.

All readings will be made available on Quercus. We will read several chapters from Michael Kearns and Aaron Roth's book, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (Oxford University Press, 2020). While these chapters will also be provided on Quercus, you may wish to acquire a hard copy.

2 Learning Outcomes

- Gain familiarity with the three core areas of AI ethics: fairness, privacy, and explainability.
- Acquire the skills necessary (i) to ethically assess applications of emerging technology, such as AI, big data, and machine learning, and (ii) to comprehend and engage with ongoing scholarship on these issues.
- Enrich one's capacity for critical thinking in general by (i) carefully *reading* scholarly literature, (ii) *discussing* this literature with others in a collaborative endeavour to understand, (iii) *writing* exegetically and critically about this literature, and (iv) *presenting* on a pertinent standpoint within it or an application of it.

3 Requirements

1. Attendance & participation. **20%**.
2. In-class essay quizzes. (3 x 13 $\frac{1}{3}$ %) **40%**.
3. Group presentation. **40%**.

Attendance & participation Attendance *and* participation are required. This is a text- and discussion-based course; each meeting therefore requires you to prepare beforehand by working through the assigned reading on your own. Bring the readings to class in hard copy (see

the **digital detox policy** below). During the lectures, we'll often return to the readings to work through a question or issue.

You may miss *two* classes without penalty, no questions asked. Further absences will result in a loss of 15% of your participation mark. Your mark for this requirement will be determined by the quality of your participation minus any penalty for absences.

In-class essay quizzes Throughout the course, there will be three in-class essay quizzes, each focusing on a core area of the ethics of AI. These quizzes will take place at the beginning of the class sessions on **October 8**, **October 22**, and **November 12**. Each quiz will consist of 2-3 questions, and you will have 40 minutes to complete it. The purpose of these low-stakes assessments is to provide you with an opportunity to gauge your comprehension of core area. Please also come prepared to discuss the new readings.

Group presentation Group presentations will occur in the last two class periods, **November 19** and **November 26**. In lieu of a final paper, these presentations are occasions to explore a topic within the ethics of AI that interests you. They also allow you to demonstrate your competency regarding your chosen topic during the question-and-answer period (Q&A).

Each group will consists of approximately 4 students, and will have 37.5 minutes to present, including Q&A. Accordingly, individual presentations should be **5–6 minutes long**. Groups will be formed according to students' proposed topics.

Individual presentations should have three parts: a *topic*, a *position*, and an *argument*. The topic denotes the ethical issue of interest to you and the necessary background information for understanding it; the position denotes the stance you wish to take on this issue; and the argument denotes the reason (or reasons) why your stance is the best one available. To establish a position and argument, you will need to contrast your view with at least one other position or argument. You may present on any topic in the Ethics of AI.

For example, consider a presentation defending balance as a measure of fairness *à la* Hellman (2020). This presentation would need to do the following: first, explain the topic—what is AI fairness and what are the three predominant fairness notions in the literature?; second, stake out a position—balance should be preferred as a measure of AI fairness over calibration and statistical parity, on the one hand, and substantial notions of fairness, on the other; third, defend this position—only balance represents what we ought to do about a classification, whereas calibration and other measures represent merely what we should believe about it, and this holds independently of the impossibility results.

By **October 22**, you must upload a presentation proposal to Quercus. This brief proposal should include your topic, position, and a sketch of your argument (1–2 sentences for each is sufficient). It is OK if you are unsure about your argument and position; you are permitted to change these at any time. If you wish to change your topic, this is also acceptable, but please discuss it with me first.

I will review your proposal, but it will not be graded. However, failure to submit a proposal will lower your presentation mark by 10% for each day it is outstanding, starting at 12:01 a.m. on October 23.

4 Policies

Digital detox policy At the beginning of each class, you are required to put away all of your electronic devices into your bag and leave them there—just like an exam. Think of this as a temporary digital detox. Have the course reading(s), notepad, and pen or pencil with you. *I will provide you will hard copies of the readings*.

Students with accessibility needs are exempt; students who would like to use electronic devices should contact me immediately so that we can work out an appropriate exception.

Communication The best way to contact me is in person—either before or after class or during my office hours (see below). The second best way to contact me is by email. Before emailing me, check the syllabus and all announcements. Ensure that your email contains all relevant information. I will try to respond within 48 hours. Redundant or incomplete emails will not receive a response. Do not use Quercus's message system.

Office hours Office hours are *drop-in* and *by appointment*, either online or in person, whatever works best for you. I'd love to see you at my office hours! In person office hours will take place on Tuesdays 2–2:50 at The Exchange, a cafe in the Rotman building across from Robarts Library (105 St George St). You can book office hours with me using Bookings (link). I recommend you prepare your questions beforehand so we can make the best use of our time.

Missed quizzes Students who miss a quiz will receive a zero. There are no make-up quizzes. It is imperative that you plan accordingly.

Acknowledgment of land We can never work to end systematic and institutional violence if we do not centre the narratives of indigenous folks in our collective decision-making for social justice and equity. As settlers in Turtle Island, we directly benefit from the colonization and genocide of the indigenous people of this land. In order to engage in resistance and solidarity against the injustices inflicted on the indigenous people of this land, it is imperative we constantly engage in acts of decolonization. Therefore, I would like to acknowledge that we are on the traditional lands of the "Mississaugas of the Credit First Nation" peoples, the traditional caretakers of this land. I would also like to pay my respects to their elders past and present, and to any who may be here with us today, physically, mentally, emotionally and spiritually. Adopted from the UTM Student Union (link).

Accessibility services Students with diverse learning needs are welcome in this course. Please contact the Accessibility Centre (link) for a needs assessment and to make arrangements.

Academic integrity All students, faculty and staff are expected to follow the University's guidelines and policies on academic integrity. For students, this means following the standards of academic honesty when writing assignments, collaborating with fellow students, and writing tests and exams. Ensure that the work you submit for grading represents your own honest efforts. Plagiarism—representing someone else's work as your own or submitting work that you have previously submitted for marks in another class or program—is a serious offence that can result in sanctions. Consult the Code of Behaviour on Academic Matters for a complete outline of the University's policy and expectations. For more information, please see Student Academic Integrity website (link).

5 Schedule

Sep 3 Foundations I — Syllabus & Introduction

- Scarfe et al., "A Real-World Test of Artificial Intelligence Infiltration of a University Examinations System" (2024)
- Suggested further reading
  - Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" (2021)

Sep 10 Foundations II — The Linear Model; What are Big Data, AI, and LLMs?

- James et al., *An Introduction to Statistical Learning* (2023), ch. 2
- Suggested further reading
  - James et al., *An Introduction to Statistical Learning* (2023), chs. 3.1 and 3.2
  - Babic et al., "Direct-to-Consumer Medical Machine Learning and Artificial Intelligence Applications" (2021), Box 1
  - Faraway, *Linear Models with R* (2014), ch. 12 [Insurance Redlining]

Sep 17 Fairness I — What is Fair AI? Technical Approaches

- Kleinberg et al., "Inherent Trade-Offs in the Fair Determination of Risk Scores" (2017), §§1 and 5, pp. 1–8, 17–18 ["Any assignment of risk scores can in principle be subject to criticisms on the grounds of bias"]
- Kearns and Roth, *The Ethical Algorithm* (2020), ch. 2, pp. 65–93 [Argues for either calibration or balance depending on the social context]
- Castro, "Just Machines" (2022) [Argues against both calibration and balance]
- Suggested further reading
  - Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments" (2017) [Impossibility results for COMPAS, clearly presented]
  - Hellman, "Measuring Algorithmic Fairness" (2020) [Defends balance]
  - Corbett-Davies et al., "The Measure and Mismeasure of Fairness" (2024) [Most comprehensive overview; introduces utility]

Sep 24 Fairness II — What is Fair AI? Philosophical Approaches

- Johnson, "Algorithmic Bias: On the Implicit Biases of Social Technology" (2021) ["There is no such thing as an unbiased program"]
- Hedden, "On Statistical Criteria of Algorithmic Fairness" (2021) [Defends calibration]

- Suggested further reading
  - Hooker, "Fairness" (2005)
  - Lippert-Rasmussen, "The Badness of Discrimination" (2006)
  - Gendler, "On the Epistemic Costs of Implicit Bias" (2011)
  - Babic and King, "Algorithmic Fairness and Resentment" (2021)
  - Creel and Hellman, "The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems" (2022)
  - Johnson, "Varieties of Bias" (2024)

Oct 1 Fairness III — Fairness Case Study

- Pro Publica, "Machine Bias" (2016)
- Pro Publica, "How We Analyzed the COMPAS Recidivism Algorithm" (2016)
- Corbett-Davies et al., "A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It's Actually Not That Clear" (2016)
- Suggested further reading
  - Dieterich et al., "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity" (2016) [Defense of COMPAS by Northpointe]
  - Flores et al., "False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks,"' (2016) [Defense of COMPAS by criminologists]

Oct 8 Data & Privacy I

- **Essay Quiz #1**
- Mattu and Hill, "How a Company You've Never Heard of Sends You Letters about Your Medical Condition" (2017)
- Crawford, *Atlas of AI* (2021), ch. 3
- Suggested further reading
  - Roessler and DeCew, "Privacy" (2023), §3.3 [Definition of access and control views of privacy]
  - Gebru et al., "Datasheets for Datasets" (2021) [An influential call-to-action for more ethical AI practice]
  - Barocas and Selbst, "Big Data's Disparate Impact" (2016) [An extensive overview of legal issues related to data]

Oct 15 Data & Privacy II

- Kearns and Roth, *The Ethical Algorithm* (2020), ch. 1
- Nissenbaum, "A Contextual Approach to Privacy Online" (2011)
- Suggested further reading
  - Dwork and Roth, *The Algorithmic Foundations of Differential Privacy* (2014), chs. 1 and 2, pp. 5–24

Oct 22 Explainability I — What is Explainable AI?

- **Essay Quiz #2**
- **Upload Presentation Proposal to Quercus**
- Creel, "Transparency in Complex Computational Systems" (2020)
- Sullivan, "Understanding from Machine Learning Models" (2022)
- Suggested further reading
  - Mittelstadt, "Interpretability and Transparency in Artificial Intelligence" (2023) [AI Interpretability contrasted with AI transparency]
  - Ribeiro et al., "'Why Should I Trust You?"': Explaining the Predictions of Any Classifier," (2016) [LIME algorithm for explaining AI, increasing functional transparency, according to Creel]
  - Mordvintsev et al., (2015) "Inceptionism: Going Deeper into Neural Networks" [Visualization for explaining AI, increasing run transparency, according to Creel]
  - Wang and Kosinski, "Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images" (2018) [Example of link uncertainty, according to Sullivan]

Reading Week

Nov 5 Explainability II — Challenges for Explainable AI

- **Meet with Presentation Groups**
- Kearns and Roth, *The Ethical Algorithm* (2020), ch. 5, pp. 169–75
- Babic and Cohen, "The Algorithmic Explainability 'Bait and Switch"' (2023)
- Suggested further reading
  - Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" (2019)

Nov 12 Blind Spots in AI Ethics?

- **Essay Quiz #3**
- **Meet with Presentation Groups**
- Hagendorff, "Blind Spots in AI Ethics" (2022), §1 (3pp)
- Castro et al., "Does Predictive Sentencing Make Sense?" (2024), §§1–2, 6–7 (6pp)

Nov 19 Presentations I

- Group presentations 1–4

Nov 26 Presentations II

- Group presentations 5–8

Exam Period