COS 598I (Spring 2024): Responsible AI in Societal Deployment

Meeting times:

Wednesday/Friday 1.30-2.50 pm. Location: Robertson 001

Course Description:

What responsibilities, and for whom? This graduate-level course addresses the theoretical and sociotechnical foundations of responsible artificial intelligence (AI) for societal deployment. Throughout the course, we will survey current methodological approaches, as well as the philosophical, ethical, and legal principles that underlie the human decisions and judgments involved in operationalizing responsible AI. Methodological aspects include: fundamentals of data-driven decision-making, choices related to data and problem formulation, limitations of prediction modeling, strategies for interventions and causal inference, techniques for measuring and mitigating algorithmic biases, uses and limitations of algorithmic fairness, dynamics of human-AI interaction, and evaluation of downstream impacts through experimental design. We will also place emphasis on understanding the broader impacts, including social justice and socioeconomic welfare, of data-driven decision and inference technologies. This will be illustrated through case studies from critical domains such as education, employment, healthcare, finance, criminal justice, and social services. Students will reflect on the values expressed in their work and complete a significant empirical research project.

Prerequisites:

The course is open to all **graduate and undergraduate students**, although extensive preparation in machine learning / artificial intelligence is expected and required (COS 324 or equivalent).

Instructor:

Prof. Lydia Liu *(she/her)*
ltliu (at) princeton (dot) edu
Office Hours: Fridays, 3 - 4 pm
Location: CS 414

Teaching Assistant:

Varun Rao *(he/him)*
varunrao (at) princeton (dot) edu

Office Hours: Wednesdays, 12.30 pm - 1.30 pm
Location: Sherrerd Hall 3rd Floor Common Area (CITP)

Gradescope and Ed

We will be using Gradescope for submitting assignments and Ed for questions and discussions. Both can be accessed via Canvas. All course materials (readings, slides) are uploaded to Ed resources. Communication with instructors should take place primarily through Ed; **please reserve email for emergencies only.**

Reading List

The following books are recommended (optional) references. Many of these are accessible for free digitally, or via the Princeton Library.

- Barocas S., Hardt M., and Narayanan A. (2023). **Fairness and Machine Learning: Limitations and Opportunities.** MIT Press.
- D'ignazio, C., & Klein, L. F. (2023). **Data feminism.** MIT press.
- Hernán MA, Robins JM (2020). **Causal Inference: What If.** Boca Raton: Chapman & Hall/CRC.
- Pearl, J., Glymour, M., & Jewell, N. P. (2016). **Causal inference in statistics: A primer.** John Wiley & Sons.
- Rawls, J. (2020). **A theory of justice: Revised edition**. Harvard university press.

Class presentations and discussions

There will be **2 student presenters** and **1 discussant** in every 80 minute class (with a few exceptions on the schedule). These will be 20-25 minute talks with the goal of (1) practicing the skill of reading and presenting the highlights of a paper (2) setting the stage for class discussion. Each student presentation will be followed by 5 minutes of Q&A, where the rest of the class can ask questions to clarify understanding of the paper's content. More resources on how to give a talk.

After the presentations, the discussant has approximately 10 minutes to reflect on the two presented papers (strength, weaknesses, and any synergies they share). The role of the discussant is to introduce a critical perspective and an attempt at synthesis. A typical presentation by a discussant would include – as a general guideline – a title page and no more than five slides. One of these slides should pose thoughtful questions to simulate conversation with the rest of the class. Further guidance on the discussant role.

Schedule of Course Meetings

Presenters and Discussants schedule

Week 1: Monday, January 29 - Friday, February 2, 2024

Day 1 Introduction and Course Overview

- Barocas, Hardt and Narayanan [abbreviated BHN], Chapter 1 (Introduction)
- IN CLASS REFLECTION: Why am I here? What are my "keystone problems"?

Day 2 Machine learning life cycle

- Liu, Wang, Britton and Abebe (2023). Reimagining the Machine Learning Life Cycle to Improve Educational Outcomes of Students.
- Suresh and Guttag (2021). Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle

Additional readings

- Chen et al. (2021). Ethical Machine Learning in Healthcare

Week 2: Monday, February 5-Friday, February 9, 2024

Day 1 Problem Selection and Problem Formulation

- D'ignazio & Klein [abbreviated DK], Introduction (optional), Chapter 1, 2.
- Obermeyer, Ziad, et al. "Dissecting racial bias in an algorithm used to manage the health of populations." Science 366.6464 (2019): 447-453.

Day 2 High-Stakes Automated Decision Making

- BHN, Chapter 2 (When is automated decision making legitimate?)
- Wang, Angelina, et al. "Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy." Available at SSRN (2022).

Additional readings:

- Saxena, Devansh, et al. "A framework of high-stakes algorithmic decision-making for the public sector developed through a case study of child-welfare." Proceedings of the ACM on Human-Computer Interaction 5.CSCW2 (2021): 1-41.

Friday, February 9 is the Undergraduate Deadline to Add/Drop Courses

Week 3: Monday, February 12 - Friday, February 16, 2024

Day 1 — Datasets in Machine Learning Research
- BHN, Chapter 9 (Datasets)
- Paullada, Amandalynne, et al. "Data and its (dis) contents: A survey of dataset development and use in machine learning research." Patterns 2.11 (2021).

Day 2 — Data in Societal level AI deployments
- Katherine Lee, Daphne Ippolito, and A. Feder Cooper. The Devil is in the Training Data
- Birhane, Abeba, et al. "Into the laions den: Investigating hate in multimodal datasets." arXiv preprint arXiv:2311.03449(2023).
    - Talk (watch before class)

Additional readings:

- Day 1 topic
    - Ding, Frances, et al. "Retiring adult: New datasets for fair machine learning." Advances in neural information processing systems 34 (2021): 6478-6490.
    - K Peng, A Mathur, A Narayanan (2021). Mitigating dataset harms requires stewardship: Lessons from 1000 papers (Dataset life cycle)
- Day 2 topic
    - Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. "Multimodal datasets: misogyny, pornography, and malignant stereotypes." arXiv preprint arXiv:2110.01963(2021).
    - P. Samuelson, "Generative AI meets copyright," Science, vol. 381, no. 6654, pp. 158–161, 2023.

Week 4: Monday, February 19 - Friday, February 23, 2024

Day 1 [2/21] — Algorithmic bias and fairness metrics
- BHN, Chapter 3 (Classification)

Day 2 [2/23] — Algorithmic fairness deep dive: Calibration
- Quiñonero Candela, Joaquin, et al. "Disentangling and Operationalizing AI Fairness at LinkedIn." Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023.
- Barda, Noam, et al. "Addressing bias in prediction models by improving subpopulation calibration." Journal of the American Medical Informatics Association 28.3 (2021): 549-558.

Additional readings
- BHN, Chapter 4 (Relative notions of fairness)

- Liu, Lydia T., Max Simchowitz, and Moritz Hardt. "The implicit fairness criterion of unconstrained learning." International Conference on Machine Learning. PMLR, 2019.
- Hébert-Johnson, Ursula, et al. "Multicalibration: Calibration for the (computationally-identifiable) masses." International Conference on Machine Learning. PMLR, 2018.

Week 5: Monday, February 26- Friday, March 1, 2024

Day 1 — Algorithmic fairness and causality
- BHN, Chapter 5 (Causality)

Day 2 — Algorithmic fairness and causality - case study
- Zanger-Tishler, Michael, Julian Nyarko, and Sharad Goel. "Risk scores, label bias, and everything but the kitchen sink." *arXiv preprint arXiv:2305.12638* (2023). **[Guest talk]**
- (Pre-reading) Richardson, Rashida, Jason M. Schultz, and Kate Crawford. "Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice." NYUL Rev. Online 94 (2019): 15.
  - Talk (watch before class)

Additional readings
- Pearl, Glymour, Jewell, Chapter 1
- Hu, Lily, and Issa Kohler-Hausmann. "What's sex got to do with machine learning?." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020.
- Kasirzadeh, Atoosa, and Andrew Smart. "The use and misuse of counterfactuals in ethical machine learning." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.

Week 6: Monday, March 4 - Friday, March 8, 2024

Day 1 — Beyond Accuracy: Arbitrariness from a sociotechnical perspective
- Black, Emily, Manish Raghavan, and Solon Barocas. "Model multiplicity: Opportunities, concerns, and solutions." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.

Day 2 — Algorithmic arbitrariness
- Marx, Charles, Flavio Calmon, and Berk Ustun. "Predictive multiplicity in classification." International Conference on Machine Learning. PMLR, 2020."

- Cooper, A. Feder, et al. "Arbitrariness and Prediction: The Confounding Role of Variance in Fair Classification." AAAI 2024

Additional readings

- Ganesh, Prakhar, et al. "On The Impact of Machine Learning Randomness on Group Fairness." Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 2023.
- Marx, Charles, Flavio Calmon, and Berk Ustun. "Predictive multiplicity in classification." International Conference on Machine Learning. PMLR, 2020."
- Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." Statistical science 16.3 (2001): 199-231.
- Semenova, Lesia, Cynthia Rudin, and Ronald Parr. "On the existence of simpler machine learning models." Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.
- Semenova, Lesia, et al. "A Path to Simpler Models Starts With Noise." Advances in Neural Information Processing Systems 36 (2024).

******Spring Recess begins at the end of classes on Friday, March 8 and runs through Sunday, March 17, 2024******

Week 7: Monday, March 18 - Friday, March 22, 2024

Day 1	Post-deployment evaluation: from impact modeling to causal experiments
Day 2	Experimental design case studies

Readings:

- Liu, Lydia T., et al. "Delayed impact of fair machine learning." International Conference on Machine Learning. PMLR, 2018.
- Chapter 1 and 2, Hernán MA, Robins JM (2020). Causal Inference: What If [preprint]
- Imai, Kosuke, et al. "Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment." Journal of the Royal Statistical Society Series A: Statistics in Society 186.2 (2023): 167-189.

Week 8: Monday, March 25 - Friday, March 29, 2024

Day 1	Post-deployment evaluation: observational studies • Perdomo, Juan C., et al. "Difficult Lessons on Social Prediction from Wisconsin Public Schools." arXiv preprint arXiv:2304.06205 (2023).
Day 2	Impact of generalization and distribution shift

- Yang, Yuzhe, et al. "Change is hard: A closer look at subpopulation shift." arXiv preprint arXiv:2302.12254 (2023).
- Pfohl, Stephen R., et al. "A comparison of approaches to improve worst-case predictive model performance over patient subpopulations." Scientific reports 12.1 (2022): 3254.

Additional readings

- Schrouff, Jessica, et al. "Diagnosing failures of fairness transfer across distribution shift in real-world medical settings." Advances in Neural Information Processing Systems 35 (2022): 19304-19318.
- Singh, Harvineet, et al. "Fairness violations and mitigation under covariate shift." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.
- DeGrave, Alex J., Joseph D. Janizek, and Su-In Lee. "AI for radiographic COVID-19 detection selects shortcuts over signal." Nature Machine Intelligence 3.7 (2021): 610-619.

Week 9: Monday, April 1 - Friday, April 5, 2024

Day 1 — Auditing - social media & employment/political ads

- Ali, Muhammad, et al. "Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes." Proceedings of the ACM on human-computer interaction. CSCW (2019): 1-30.

Day 2 — "Audits & Accountability in the Age of Artificial Intelligence" (William Pierson Field Lecture, Deborah Raji)

- Raji, Inioluwa Deborah, et al. "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing." *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020.
- Raji, Inioluwa Deborah, et al. "Outsider oversight: Designing a third party audit ecosystem for ai governance." *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 2022.

Additional readings

- Ad Delivery Algorithms: The Hidden Arbiters of Political Messaging: https://dl.acm.org/doi/pdf/10.1145/3437963.3441801
  - political ads delivered to users perceived to be politically aligned to ads messaging
- Auditing for Discrimination in Algorithms Delivering Job Ads: https://dl.acm.org/doi/pdf/10.1145/3442381.3450077

    - skew in job ad delivery cannot be explained due to differences in qualification
- Measurement and Analysis of Implied Identity in Ad Delivery Optimization: https://dl.acm.org/doi/pdf/10.1145/3517745.3561450
    - Ads are often delivered disproportionately to users similar to those pictured
- Discrimination through Image Selection by Job Advertisers on Facebook: https://dl.acm.org/doi/pdf/10.1145/3593013.3594115
    - real world evidence of job advertisers using images of selective demographic

Week 10: Monday, April 8 - Friday, April 12, 2024

Day 1	Human stakeholders and operators ● Lee, Min Kyung, et al. "WeBuildAI: Participatory framework for algorithmic governance." CSCW 2019.
Day 2	Human stakeholders and operators ● Alur, Rohan, et al. "Auditing for Human Expertise." Advances in Neural Information Processing Systems 36 (2023).

Week 11: Monday, April 15 - Friday, April 19, 2024

Day 1	Economic impacts: labor
Day 2	Inequality ● BHN, Chapter 4. "Relative Notions of Fairness"

Readings:

- Zhang, Angie, et al. "Algorithmic management reimagined for workers and by workers: Centering worker well-being in gig work." CHI 2022.
- Rosenblat, Alex, and Luke Stark. "Algorithmic labor and information asymmetries: A case study of Uber's drivers." International journal of communication (2016)
- Dell'Acqua, Fabrizio, et al. "Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality." Harvard Business School Technology & Operations Mgt. Unit Working Paper 24-013 (2023).
- Bell, Alex, et al. "Who becomes an inventor in America? The importance of exposure to innovation." The Quarterly Journal of Economics 134.2 (2019): 647-713.
- Mary L. Gray and Siddharth Suri. **Ghost Work**. CHAPTER 1, 2.
- Rawls, J. (2020). **A theory of justice: Revised edition**. CHAPTER 1.

- Dubal, Veena. "On algorithmic wage discrimination." Columbia Law Review 123.7 (2023): 1929-1992.

Week 12: Monday, April 22 - Friday, April 26, 2024

Day 1	Machine values and ethics: Social Justice, Democracy, Power ● Cohen and Liu manuscript
Day 2	Final Project Presentations

Additional readings:

- Himmelreich, Johannes. "Ethics of technology needs more political philosophy." Communications of the ACM 63.1 (2019): 33-35.
- "Justice as Fairness: Justice within a Liberal Society" Stanford Encyclopedia of Philosophy

****Friday, April 26, 2024 is the last day of scheduled course meetings for the spring term in 2024****

Grading and assessment requirements

Percentage of final grade / %	**Coursework**
20	Final oral presentation
20	Final project report
40	Assignments (3 individual assignments and 1 group project proposal)
20	Class participation

The success of this course depends on your commitment to complete all required readings for each class meeting, to critically reflect on the readings, to sign up to present course material (when there is an opportunity), to participate actively in class discussions, and to creatively integrate these insights into the final project. Full attendance and active participation in class discussions are expected.

Due dates

- HW1 is due in February (date TBD)
- HW2 is due before spring break (date TBD)
- HW3 is due in April (date TBD)
- Group project proposal is due after spring break (date TBD)
- Final project report is due on Dean's Date May 7, 2024

Lateness and extensions

You are given 1 late pass for any course work (no late pass for the final project). A late pass allows you to turn in an assignment up to 24 hours late without a penalty. **The late pass cannot be used for the final project.** The first time you are late for submitting an assignment, the late pass will automatically be applied. Note: **Do not email any of the instructors to just let them know that you are using a late pass.**

Unless otherwise stated, all deadlines in this course syllabus are firm. **Work that is not submitted on time will be subject to a 20% penalty for each day it is late**, unless a late pass is applied. No other extensions will be granted. Homework that is more than 3 days late will not be graded and will receive 0 credit. Please do not email any of the instructors to ask for an extension.

Disability Services and Academic Accommodations

Princeton welcomes students with disabilities and values their diverse experiences and perspectives. If you anticipate or experience a barrier to learning in the classroom or in completing assignments or exams, please know there is support for you. Students who wish to request classroom accommodations can do so through the Office of Disability Services (ODS). If you have been approved for accommodations through the ODS, please contact me via email as soon as possible so we can develop an implementation plan together.

Academic Integrity and the Honor Code

Intellectual honesty is vital to an academic community and for my fair evaluation of your work. For these reasons, all students in this course are expected to abide by the Honor Code on examinations and to complete their written work in accordance with University regulations. All work submitted in this course must be your own, completed in accordance with the University's academic regulations. Unless otherwise stated, you are welcome to use AI tools such as code generators and text generators for written work. Of course, you are responsible for the accuracy and quality of the work that you submit. If you decide to use ChatGPT to support the work on

your course assignments in any way, you **must** include an acknowledgment to that effect and, as part of your submission, briefly explain how you leveraged the tool and what you think about the effect on the quality of your work.

Recording Policy and Electronic Devices

Unauthorized photographic, video and/or audio recording of class activities (lectures, workshops, precepts, exercises) is prohibited. Students eligible for approved accommodations through the Office of Disability Resources due to religious observance or other reasons approved by the Office of the Dean of the College are strongly encouraged to alert the professor as soon as possible so that appropriate arrangements can be made. On occasion, the instructor may grant permission for individual students to record particular course activities; in such instances, the professor will make every effort to notify all course contributors in advance that class activities will be recorded. Such recordings are only to be used to directly support active participation in the work of the course during the term of enrollment by the student receiving the accommodation. Further, such recordings are not to be shared, distributed, or retained beyond the semester's end. Disregard for the specific terms of accommodation allowing the recording of course activities will be considered a violation of academic integrity.

Commitment to Diversity, Equity and Inclusion

It is essential we build our class community into a place where everyone feels comfortable participating. Disrespect or discrimination on any basis will not be tolerated. We will strive to create a learning environment that supports a diversity of thoughts, perspectives, and experiences. All members of this class are expected to contribute to a respectful, welcoming, and inclusive environment for every other member of the class.

Preferred names and pronouns

Instructors are provided with class rosters with potentially out-of-date preferred names; information about students' pronouns is also not readily available to us. We encourage any corrections and look forward to learning your pronouns.

Land Acknowledgement

This course will be facilitated from Princeton, New Jersey – or the unceded, ancestral land of the Lenni-Lenape. As we gather, we honor the ongoing history and living culture of the Nanticoke Lenni-Lenape people, other Indigenous caretakers of these lands and waters, the elders who lived here before, the Indigenous people living today both in and beyond this space

and the generations yet to come. For information about the histories of Indigenous stewardship of the land on which you reside, consult Native Land Digital.

Mental Health Resources

Princeton University offers a variety of resources to support your mental health and wellbeing. If you or someone you know needs support or is looking to access specific services, consider reaching out to these university and student-led resources:

- Your residential college advising team is always a good first resource for advice and counsel. The assistant deans for student life (DSLs), whose offices are located in each residential college, serve as case managers in crisis situations. They are also available to talk with you about well-being concerns and can refer you to appropriate campus resources.
- If you are feeling distressed or need support, please contact Counseling & Psychological Services (CPS) at 609-258-3141 for immediate support or to schedule an appointment with a counselor. CPS is a confidential resource.
- The Sexual Harassment/Assault Advising, Resources and Education (SHARE) office is a survivor-centered, trauma-informed, confidential resource on campus. SHARE provides crisis response, support, counseling, advocacy, education, and referral services to students experiencing unhealthy relationships and abuse, including harassment, sexual assault, dating/domestic violence, and stalking.
- The Princeton Peer Nightline is a student-run anonymous peer listening service. It is not affiliated with CPS or the University administration. They offer anonymous chat/call peer support.

It is your responsibility to balance your personal commitments so that you can also meet your curricular obligations in this class. However, if a personal situation arises during the semester that threatens to affect your performance in the course, you are strongly encouraged to speak with the professor and your residential college dean or assistant dean for student life about your circumstances immediately.