Fall 2020 Syllabus for CSE 544A Special Topics in Artificial Intelligence

Equity and Fairness in Estimation and Classification

Instructor: Neal Patwari, Professor in ESE and CSE, npatwari@wustl.edu

Course consulting: Credit to Dr. Cathleen Power for consulting in the design of course and its readings

Course website: Canvas

Meetings: Tue/Thu 2:30-3:50pm Central time

Prerequisites: ESE 326 (Probability and Statistics) and CSE 417T (Machine Learning), or equivalent; background from critical disciplines. Background readings will be available.

Required Reading: Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code*, July 2019, ISBN: 978-1-509-52643-7.

Other Readings: Readings include excerpts from books, the recent manuscripts and publications from conferences (FAccT, NeurIPS, arXiv), videos from tutorials and presentations, journal papers and book chapters.

Course objective: This course is designed to introduce graduate engineering students to research at the intersection of engineered algorithms (such as estimators, detectors, classifiers, and control systems) and the problems of bias, unfairness, inequity, and oppression. The objective is both to use engineering tools to analyze and quantify problems, as well as to study the problems introduced by engineering tools including detectors, estimators, classifiers, and control systems. Topics include critical race and feminist theory, measurement theory, estimation bias, limitations on detection fairness, and random process models used for control systems.

Topics:

- Introduction to critical science: systemic oppression, power, bias, (in)equity, and (un)fairness
- Measurement theory
- Fairness issues in detection and limitations on fair detectors
- Meanings of estimation “bias”, study of word embeddings
- Studying up: use of data science to confront power systems
- Feedback and game theory in systems with control loops; random process models with reinforcement

Grading: This course will be graded based on participation, reflection writing, quizzes, class leading, and student semester project. Students will present a project on a research topic of their choosing, which will involve leading a discussion, and a written report.

Meetings: This course will meet twice per week on Zoom (Tue/Thu 2:30-3:50pm CT). Meetings will be used both for lecture and for discussion; it is critical that students are prepared to participate in a **synchronous discussion** during the meeting time.

Learning is Co-Creation: I note that our effort to achieve the above learning goals will be a joint effort. True understanding and progress towards the learning goals will require your active participation as both learner and contributor. Sometimes the goal of learning and co-creation of knowledge is hampered by the goal of grading. For example, taking risks is necessary to speak up, to generate new knowledge, but grading can make the risk not worth it. To work to avoid this, we will at the start of the semester collaboratively build our assignments and grading system that represents the consensus of the class that operates with the constraint of ensuring that the learning goals are met. We will also be open to altering the schedule of readings/media in order to better achieve our learning goals.

Course Schedule:

Course Introduction and Construction

Readings (Media) by Class
● We will discuss in class the syllabus. We will discuss dialog via the handout by Ratnesh Nagda, Patricia Gurin, Jaclyn Rodriguez & Kelly Maxwell (2008).
First showing of *Coded Bias* at 6:30pm on Tue Sept 15. Pre-register.
● Jay Smooth, How I Learned to Stop Worrying and Love Discussing Race, TEDx talk, Nov 15, 2011. He references his viral video, How To Tell Someone They Sound Racist ● Paulo Freire, Pedagogy of the Oppressed. Read Chapter 2.
Second showing of *Coded Bias* at 3pm on Sun Sept 20. Pre-register
● Buolamwini, Joy, and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In ACM Conference on Fairness, Accountability and Transparency (FAccT 2018), pp. 77-91, 2018. Paper website: http://gendershades.org/overview.html ● *Coded Bias*, the documentary film which you have watched via the online film festival (above)

Critical Theory

Background Reading:

- *Social Identities, Power, and Oppression*: National Museum of African American History & Culture, "Talking About Race: Social Identities and Systems of Oppression", Web resource, https://nmaahc.si.edu/learn/talking-about-race/topics/social-identities-and-systems-oppression.
- *Outcomes are inequitable by race in St. Louis*: City of St. Louis, "Equity Indicators Baseline Report: City of St. Louis Equity Indicators Baseline 2018 Report", Jan 2 2019. https://www.stlouis-mo.gov/government/departments/mayor/initiatives/resilience/equity/documents/equity-indicators-baseline-report.cfm.
- Race in the US: The 1619 Project. Nikole Hannah-Jones, https://www.nytimes.com/1619, *The New York Times Magazine*, 2019.
- Criminal Justice in the US: 13th, Netflix documentary film, Released October 7, 2016.

Readings (Media) by Class
● Neal Patwari, Whose Tools? There's a reason we narrow our set of tools in CS & engineering, Aug 29, 2020. ● Marilyn Frye, Oppression, in The Politics of Reality: Essays in Feminist Theory, The Crossing Press, 1983, pp. 1-16. ● Audre Lorde, The Uses of Anger, Women's Studies Quarterly , Spring - Summer, 1997, Vol. 25, No. 1/2, Looking Back, Moving Forward: 25 Years of Women's Studies History (Spring - Summer, 1997), pp. 278-285. Speech in Fall 1981.
● Peggy McIntosh, "White Privilege: Unpacking the Invisible Knapsack", Peace and Freedom Magazine, July/Aug 1989, pp. 10-12. ● Gina Crosley-Corcoran, "Explaining White Privilege To A Broke White Person", Huffington Post, 05/08/2014 12:57 pm ET, Updated Dec 06, 2017.
● Take an IAT at Project Implicit, https://implicit.harvard.edu/implicit/takeatest.html, read the About and FAQ ● Ibram X. Kendi, How to be an Anti-Racist, Chapter 1, Chapter 3. The WashU library has an ebook, but only 3 licenses. Please plan ahead so that <=3 of you are reading at the same time.
● Hanna, Alex, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT 2020), pp. 501-512. 2020. Paper presentation: https://www.youtube.com/watch?v=aW-6Kq_tZv0

● Keyes, Os. The misgendering machines: Trans/HCI implications of automatic gender recognition. Proceedings of the ACM on Human-Computer Interaction 2, no. CSCW (2018): 1-22. ● Caterina Fake, host. "Affectiva: Software That Detects How You Feel, Rana El Kaliouby, Co-founder And CEO", *Should This Exist?,* Podcast.
● Abigail Jacobs, Su Lin Blodgett, Solon Barocas, Hal Daumé III, Hanna Wallach, "The Meaning and Measurement of Bias: Lessons from Natural Language Processing", Translation Tutorial, ACM Conference on Fairness, Accountability, and Transparency (FAccT 2020), Jan 2020. Watch the video of the tutorial. ● Abigail Z. Jacobs, Hanna Wallach, "Measurement and Fairness", arXiv:1912.05511 [cs.CY].

Bias in Estimation and Classification Algorithms

Readings (Media) by Class
● Cynthia Dwork, “What’s Fair?”, Keynote at KDD 2017, Online video: ● Ruha Benjamin, *Race after Technology*. Read the book and watch her keynote talk at the *2020 Human Impacts of AI Symposium.*
● Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635 (2019). **You may skip Section 5**. ● Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD Intl. Conf on Knowledge Discovery and Data Mining*, pp. 259-268. 2015.
● Selbst, Andrew D., danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59-68. 2019.

Example System Applications

Readings (Media) by Class
● *Background Reading*: Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013). ● Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai, “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”, NeurIPS 2016.

- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of 'Bias' in NLP. Annual Meeting of the Association for Computational Linguistics (ACL). July 2020.

- Chelsea Barabas, Colin Doyle, JB Rubinovitz, Karthik Dinakar, Studying up: reorienting the study of algorithmic fairness around issues of power, FAccT 2020.
- Example of Studying Up: John Wiseman, "What's flying above us? I'm working to make it easy to find out.", website: https://skycircl.es/. Source: https://gitlab.com/jjwiseman/advisory-circular/

- Inioluwa Deborah Raji, Joy Buolamwini, Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products, in *Proc. of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Pages 429-435, 2019.

- Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, Suresh Venkatasubramanian (2018). Runaway Feedback Loops in Predictive Policing. FAccT 2018. Video of presentation.
- Technical background for the "Runaway" paper: Robin Pemantle. A survey of random processes with reinforcement. Probab. Surveys, 4:1–79, 2007. doi: 10.1214/07-PS094.
- Background on predictive policing: Kristian Lum and William Isaac, To predict and serve?. *Significance Magazine*, Oct 2016.

Classes Led by Student Groups

Class Sessions 20-26 (7 sessions in total) are led by individual groups.

- Washington, Anne L., and Rachel Kuo. "Whose side are ethics codes on? power, responsibility and the social good." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 230-240. 2020. (Link to paper) (Link to video of presentation)

- Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. "Mitigating bias in algorithmic hiring: Evaluating claims and practices." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 469-481. 2020. https://arxiv.org/pdf/1906.09208.pdf (Conference presentation video.)
- Optional reading: Cowgill, Bo, Fabrizio Dell'Acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. "Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics", In Proceedings of the 21st ACM Conference on Economics and Computation (pp. 679-681). (Conference presentation video.)

- Lee, Michelle Seng Ah, and Luciano Floridi. "Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs." Available at SSRN (2020).Link
- Flippen, C. (2004). Unequal returns to housing investments? A study of real housing appreciation among black, white, and Hispanic households. *Social Forces*, *82*(4), 1523-1551.

● One optional reading: Andre Perry, Jonathan Rothwell, David Harshbarger, "The devaluation of assets in Black neighborhoods: The case of residential property", Gallup Metropolitan Policy Program, Nov 2018.
● One required reading: Amanda R. Kube, Sanmay Das, and Patrick J. Fowler, "Fair and Efficient Allocation of Scarce Resources Based on Predicted Outcomes: Implications for Homeless Service Delivery", Nov 2019.
● Rashida Richardson, Jason M. Schultz, and Kate Crawford, Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. ● Michael Veale and Reuben Binns, Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data ● Optional reading: Tim Lau, "Predictive Policing Explained", Report from the Brennan Center for Justice, April 1, 2020.
● Jared Moore. 2020. Towards a more representative politics in the ethics of computer science. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 414–424. DOI:https://doi.org/10.1145/3351095.3372854 ● Explore other class sites from other universities similar to our class: 1) Computer Ethics, University of Washington; 2) Ethics and Policy in Data Science, Cornell
● Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, Wagner Meira Jr, Auditing Radicalization Pathways on Youtube, ACM FAccT 2020. ● Sirui Yao, Bert Huang, Beyond Parity: Fairness Objectives for Collaborative Filtering, 31st Conference on Neural Information Processing Systems (NIPS 2017).