Fairness, Equity, and Justice in AI and Computing

Course Syllabus for INFO 873

AT A GLANCE

Course

Course Title	INFO 873: Fairness, Equity, and Justice in AI and Computing
Credits	1
Schedule	Thu 3:00–3:50 PM
Office hours	By appointment (schedule here)
Readings	Posted to Blackboard

REVISION LOG

Jan. 30 (v2)	Finished reading list. Revised Week 9 readings.
Jan. 30	Finished readings through Week 6.
Jan. 8	Finished topic list and added most of the readings.
Jan. 6	Added notes on readings and reading responses.

COURSE INFORMATION

Course Description

Over the last 10+ years, questions of *fairness* and related concerns have emerged as an area of significant interest across many areas of computer and information science, including AI, data science, information retrieval, NLP, and human-computer interaction. Numerous definitions, metrics, evaluation methods, and intervention strategies have been proposed to define, observe, and improve fairness. Fairness also interacts in complex ways with a number of other goals, including equity, justice, transparency, and accountability. This seminar provides an introduction to research on this topic through a combination of foundational and recent literature, with the aim of providing students with a solid foundation to engage with and conduct research on fairness and equity in any of the specialties in our department.

Course Purpose Within a Program of Study

This is an elective seminar course for Ph.D. students.

COURSE STRUCTURE

This course is a **reading and discussion** course, not lecture-based. Each week (including the first) has selected readings you need to complete **before class**. Prior to class, submit a 2-5 paragraph essay responding to the readings to that week's Blackboard board (the *reading response*), and read your fellow students' responses.

During class, we will have 1-2 short (~10 minutes) presentations on selected readings, followed by discussion of the papers and the themes and findings from them. You will volunteer for presentation times and topics the first week of class.

This course is intended for students with both quantitative and qualitative emphases, and we will be reading papers from multiple methodological perspectives. We will discuss more about how different perspectives relate and inform each other in the study of AI ethics and equity in our first class meeting.

A Note on Reading

There is quite a bit of reading in this class, and we start off with reading the first week. Not every reading is required for every week; some weeks may have supplemental readings or a choice of readings. You are most likely to succeed if you read strategically, rather than trying to make sure you understand every detail of each section or paragraph before moving on. Some questions to think about to help you prioritize how you read the paper:

- What is the core goal and/or claim of the paper?
- What arguments or evidence do the authors bring in support of their claim?
- If it is a research inquiry, what methods, data sets, study sites, etc. do they use for their research?
- What do you learn from the paper to inform your understanding and/or future work?

ASSESSMENT PLAN

Grading in this course will be based on in-class presentations (60%), reading responses (20%), and class presentations (20%).

Reading Responses

As noted above, the reading responses in this class are short (2–5 paragraphs) responses to or reflections on the readings. Your response can discuss things you found particularly insightful, things you disagree with, examples you have seen in your own experience or work that either demonstrate or complicate the paper's ideas, etc.; the specific topic is up to you. I am looking to see that you are engaging with and understanding the readings, and to see what you think about the different papers we read and ideas we encounter.

COURSE POLICIES

Announcements

I will post course announcements to Blackboard Announcements and to Discord, including any changes to the syllabus or assignments. You are responsible for making sure that you receive course announcements in a timely fashion. If I need to change the syllabus or an assignment description after its initial publication, I will include a dated Revision Log in that document describing the modifications.

Late Work

Due to the discussion-based nature of this course, late work is not accepted. However, I will excuse 2 missing reading responses throughout the week.

Conduct

I expect you to respect me and your fellow students in all class interactions, both in official meetings such as lectures and out-of-classroom activities such as project group meetings and study sessions, and to contribute to a constructive learning environment.

In addition to the Drexel Conduct and Community Standards, the Recurse Center Social Rules are a good source of guidance on how to maintain a constructive and educational environment in a computing learning context.

Permitted Use of Artificial Intelligence

Use of generative AI (such as ChatGPT or CoPilot) is not allowed for the course submissions, and while AI summarizers may help you navigate the papers, they are not a substitute for reading the papers. The value of the reading reflections and presentations is from the deliverables themselves, but in the process of reading, understanding, and reflecting. Short-circuiting this process with AI will undermine your ability to learn the material and the modes of thought in this course. The Drexel Policy on Academic Integrity Pertaining to Artificial Intelligence provides further details on university policies regarding AI.

Disability Accommodations

If you need particular accommodations to be able to fully participate in this course, please talk with me as soon as possible. If you have documentation from Disability Resources for particular accommodations, please bring it, but I am happy to discuss with you anything needed for you to fully participate in the class.

Office Hours

My office hours are by appointment, arranged through my Bookings page. There is a link to this page at the beginning of the syllabus and in Blackboard.

Course Changes

I may need to make changes to the course as the term progresses to better support your learning and the logistics of delivering the course. Such changes will be announced through Blackboard and Discord, as well as mentioned in lecture when the timing of the change permits.

SCHEDULE AND READINGS

The following is our planned schedule of topics and readings. Exact readings may be adjusted or supplemented as the class progresses (especially for later weeks), and I encourage students presenting to locate additional readings to share with the class on their topics.

Week 1 (Jan. 9): Questions of Equity

Barocas, Solon, and Andrew D Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104 (3): 671. https://www.jstor.org/stable/24758720.

Friedman, Batya, and Helen Nissenbaum. 1996. "Bias in Computer Systems." *ACM Transactions on Information Systems* 14 (3): 330–47. https://doi.org/10.1145/230538.230561.

Week 2 (Jan. 16) — Concepts and Mathematics

Friedler, Sorelle A, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. "The (Im)Possibility of Fairness." *Commun. ACM* 64 (4): 136–43. https://doi.org/10.1145/3433949.

Mitchell, Shira, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2020. "Algorithmic Fairness: Choices, Assumptions, and Definitions." *Annual Review of Statistics and Its Application* 8 (November). https://doi.org/10.1146/annurev-statistics-042720-125902.

Dwork, Cynthia. 2017. "What's Fair?" In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1. KDD '17. https://doi.org/10.1145/3097983.3105807. Keynote address — watch video.

Week 3 (Jan. 23) — Recidivism and Release

This week we will look at the COMPAS data set and its legacy, as it is both a commonly-referenced touchpoint for algorithmic fairness discussion and a widely-used dataset.

Points and Counterpoints on COMPAS

Angwin, Julia, Surya Mattu, Jeff Larson, and Lauren Kirchner. 2016. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks." *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Dieterich, William, Christina Mendoza, and Tim Brennan. 2016. "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity." Northpointe. https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616/. (long, read for key ideas and claims, in more detail if you would like)

Angwin, Julia, and Jeff Larson. 2016. "Technical Response to Northpointe." *ProPublica*, July 29, 2016. https://www.propublica.org/article/technical-response-to-northpointe.

Further Analysis (optional, recommend for a presenter)

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic Decision Making and the Cost of Fairness." In *KDD '17*, 797–806. ACM. https://doi.org/10.1145/3097983.3098095.

Ongoing Impact

Bao, Michelle, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. "It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks." *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1 (December). https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/92cc227532d17e56e07902b254dfad10-Abstract-round1.html.

Additional Material

"COMPAS Recidivism Risk Score Data and Analysis - ProPublica Data Store." 2016. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis.

Week 4 (Jan. 30) — Hiring and Recruiting

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." In *FAT* '20*, 469–81. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372828.

Sánchez-Monedero, Javier, Lina Dencik, and Lilian Edwards. 2020. "What Does It Mean to 'solve' the Problem of Discrimination in Hiring? Social, Technical and Legal Perspectives from the UK on Automated Hiring Systems." In *FAT* '20*, 458–68. https://doi.org/10.1145/3351095.3372849.

Sühr, Tom, Sophie Hilgard, and Himabindu Lakkaraju. 2021. "Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring." In *AIES '21*, 989–99. https://doi.org/10.1145/3461702.3462602.

Week 5 (Feb. 6) — Search and Ranking

Biega, Asia J, Krishna P Gummadi, and Gerhard Weikum. 2018. "Equity of Attention: Amortizing Individual Fairness in Rankings." In *Proceedings of the*

41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 405–14. SIGIR '18. https://doi.org/10.1145/3209978.3210063.

Zehlike, Meike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. "FA*IR: A Fair Top-k Ranking Algorithm." In *CIKM '17*, 1569–78. https://doi.org/10.1145/3132847.3132938.

Smith, Jessie J., Lex Beattie, and Henriette Cramer. 2023. "Scoping Fairness Objectives and Identifying Fairness Metrics for Recommender Systems: The Practitioners' Perspective." In *Proceedings of the ACM Web Conference 2023*, 3648–59. https://doi.org/10.1145/3543507.3583204.

Supplementary Review Reading

Raj, Amifa, and Michael D Ekstrand. 2022. "Measuring Fairness in Ranked Results: An Analytical and Empirical Comparison." In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 726–36. https://doi.org/10.1145/3477495.3532018.

Zehlike, Meike, Ke Yang, and Julia Stoyanovich. 2022. "Fairness in Ranking, Part I: Score-Based Ranking." *ACM Comput. Surv.*, April. https://doi.org/10.1145/3533379.

Ekstrand, Michael D, Anubrata Das, Robin Burke, and Fernando Diaz. 2022. "Fairness in Information Access Systems." *Foundations and Trends® in Information Retrieval* 16 (1–2): 1–177. https://doi.org/10.1561/1500000079 [free author link].

Week 6 (Feb. 13) — Natural Language Processing (guest prof. Dr. Rezapour)

Blodgett, Su Lin, and Brendan O'Connor. 2017. "Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English." *arXiv [Cs.CY]*. http://arxiv.org/abs/1707.00061.

Narayanan Venkit, Pranav, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. "Unmasking Nationality Bias: A Study of Human Perception of Nationalities in AI-Generated Articles." In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 554–65. AIES '23. https://doi.org/10.1145/3600211.3604667.

Hovy, Dirk, and Shrimai Prabhumoye. 2021. "Five Sources of Bias in Natural Language Processing." *Language and Linguistics Compass* 15 (8): e12432. https://doi.org/10.1111/lnc3.12432.

Supplementary Reading

Dev, Sunipa, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, et al. 2022. "On Measures of Biases and Harms in NLP." arXiv. https://doi.org/10.48550/arXiv.2108.03362.

Hovy, Dirk, and Shannon L. Spruit. 2016. "The Social Impact of Natural Language Processing." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 591–98. Berlin, Germany: Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2096.

Week 7 (Feb. 20) — Stereotypes and Representational Harms

List known to be incomplete.

Raj, Amifa, and Michael D Ekstrand. 2022. "Fire Dragon and Unicorn Princess: Gender Stereotypes and Children's Products in Search Engine Responses." In *Proceedings of the 2022 SIGIR Workshop on eCommerce.* http://arxiv.org/abs/2206.13747.

Davani, Aida Mostafazadeh, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. "Hate Speech Classifiers Learn Normative Social Stereotypes." *Transactions of the Association for Computational Linguistics* 11 (March):300–319. https://doi.org/10.1162/tacl_a_00550.

Week 8 (Feb. 27) — Prediction, Feedback, and Impact

This week's class meeting will be on Zoom due to my travel.

Lum, Kristian, and William Isaac. 2016. "To Predict and Serve?" *Significance* 13 (5): 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x.

Ensign, Danielle, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. "Runaway Feedback Loops in Predictive Policing." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, edited by Sorelle A Friedler and Christo Wilson, 81:160–71. Proceedings of Machine Learning Research. New York, NY, USA: PMLR. http://proceedings.mlr.press/v81/ensign18a.html.

Epps-Darling, Avriel, Romain Takeo Bouyer, and Henriette Cramer. 2020. "Artist Gender Representation in Music Streaming." In *Proceedings of the 21st International Society for Music Information Retrieval Conference*, 248–54. ISMIR. https://program.ismir2020.net/poster_2-11.html.

Week 9 (Mar. 6) — Categories, Limitations, Critiques

Hoffmann, Anna Lauren. 2019. "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse." *Inf. Commun. Soc.* 22 (7): 900–915. https://doi.org/10.1080/1369118X.2019.1573912.

Hanna, Alex, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. "Towards a Critical Race Methodology in Algorithmic Fairness." In *FAT* '20*, 501–12. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372826.

Selbst, Andrew D, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. "Fairness and Abstraction in Sociotechnical Systems." In *FAT* '19*, 59–68. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598.

Additional Background (**recommended**)

Crenshaw, Kimberlé. 2015. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics." *University of Chicago Legal Forum* 1989 (8). https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8.

Week 10 (Mar. 13) — Co-Designing Fairness

Readings TBD.

Smith, Jessie J., Aishwarya Satwani, Robin Burke, and Casey Fiesler. 2024. "Recommend Me? Designing Fairness Metrics with Providers." In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2389–99. FAccT '24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3630106.3659044.

Schellingerhout, Roan, Francesco Barile, and Nava Tintarev. 2023. "A Co-Design Study for Multi-Stakeholder Job Recommender System Explanations." In *Explainable Artificial Intelligence*, edited by Luca Longo, 597–620. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-44067-0_30.

Venkatasubramanian, Suresh, Timnit Gebru, Ufuk Topcu, Haley Griffin, Leah Rosenbloom, and Nasim Sonboli. 2024. "Community Driven Approaches to Research in Technology & Society CCC Workshop Report." Computing Community Consortium. https://cra.org/ccc/events/community-driven-approaches-to-research-in-technology-society/.

ACKNOWLEDGEMENTS

The course and grading structure in this syllabus are based on Shadi Rezapour's seminar syllabi.