DSCI-531: Fairness in Artificial Intelligence
Units: 4
Spring 2022 MW 4:00-5:50PM

Instructor: Kristina Lerman
Office: ISI 932
Office Hours: TBD
Contact Info: lerman@isi.edu, 310-448-8714

Instructor: Fred Morstatter
Office: ISI 930
Office Hours: TBD
Contact Info: fred@isi.edu, 310-448-9381

Catalog Description

Our society's rapid algorithmification is fueled by data, but the reliance on data raises important questions. What are the latent biases hidden in the collected data? If that data was used to train machine learning algorithms, how did these biases impact predictions made by algorithms and systems that depend on them? Are the algorithmic decisions fair, or do they perpetuate stereotypes and fortify discrimination? As we come to rely on AI to make decisions in our lives and allow for synergistic relationship with technology, we need to build trust in AI by improving algorithmic fairness, accountability, transparency and explainability.

Course Description

The course will explore topics in the intersection of data, language, networks and algorithms with fairness and bias through quantitative analysis and hands-on exploration. The course will introduce students to basic and advanced fairness concepts, including methods and apply them to societal data to study fairness and bias, and to understand their effects on learning algorithms. While there are no formal prerequisites for the class, students are expected to be proficient in programming, algorithm design and data structures, and to have taken college level or above courses in linear algebra and statistics. AI and machine learning courses are strongly recommended.

Learning Objectives

After completing the course, students will be able to do the following:
- Demonstrate a familiarity with the principles of ethical considerations for AI systems. This includes definitions of fairness in computer science and related literature.
- Understand and identify sources of bias in data analysis tasks, including text, image and tabular data.
- Understand the impacts of bias on machine learning tasks.
- Choose or design bias mitigation algorithms suitable for a particular task.
- Detect and assess biases in both datasets and trained machine learning models.
- Presenting technical results via technical reports.
- Generate thought provoking peer-reviews.
- Present your findings in the form of a short presentation.

Recommended Preparation

Knowledge of Python 3; undergraduate coursework in linear algebra, basic probability and statistics; familiarity with artificial intelligence and machine learning at the level of DSCI 552.

Course Notes

The course will be run as a lecture class with student participation strongly encouraged. There are weekly readings and students are encouraged to do the readings prior to the discussion in class. All of the course materials, including the readings, lecture slides, homeworks will be posted online on Blackboard. The class project is a significant aspect of this course and at the end of the semester, students will present their projects in class.

Technological Proficiency and Hardware/Software Required

Students are expected to know how to program in a language such as Java, C++, or Python 3 (recommended). Students are also expected to have their own laptop or desktop computer where they can install and run software to do the homework assignments.

Required Readings and Supplementary Materials

There are no required textbooks. The reading material is based on recently published technical papers available via the ACM/IEEE/Springer digital libraries. All USC students have automatic access to these digital archives. Lecture slides will be placed on Blackboard and will be accessible to students before each lecture.

Description and Assessment of Assignments

<u>**Weekly Readings and Quizzes**</u>
There will be 5 random quizzes at the beginning of class. The questions will concern some basic ideas discussed in the class and/or the topics related to the recommended readings. Each quiz will be posted on Blackboard on a Monday and will be due before the start of the class the following Monday. Your lowest score will be dropped.

<u>**Homework Assignments**</u>
There will be four homework assignments designed to give students proficiency with understanding biases in different data modalities:
- **HW1: Ethics in the Research Process** - This homework will introduce students to the concepts of conducting fair and ethical research. The focus will be on historical understanding of the nature of IRB. Students will complete the CITI Human Subjects Research training as part of their assignment.
- **HW2: Bias in Data and Prediction** – Students will learn to apply basic data mining techniques to data. Students will design and conduct statistical tests on large datasets. These tests will be designed around fairness concepts, and how to leverage techniques to identify unfairness.
- **HW3: Bias in NLP --** Students will study gender bias in named entity recognition. Solving this homework requires basic natural language processing techniques including transformer-based language models like BERT.
- **HW4: Bias in Networks** -- In this homework, students will learn and apply basic network techniques to uncover the presence of gender bias in a network. Are women more or less represented in the network? Do they tend to occupy positions of higher or lower centrality than men?

Note: In both written and programming assignments, the completeness and the clarity of your description and analysis will matter as much as the final correct answer. Sending just a single final value (even if correct) is not enough. See the table below:

Grade Component	Meets Expectations (75%-100%)	Approaches Expectations (50%-75%)	Needs Improvement (0%-50%)
Completeness (50%)	All parts of the question are addressed. If the task was to a) select a machine learning algorithm, b) train, and c) validate the model, all three parts are completed.	Most parts of the question are addressed. If the task was to a) select a machine learning algorithm, b) train, and c) validate the model, the student selected and trained the model, but the validation part is missing or is incomplete.	The main question is not addressed. The answer is irrelevant to the task. The analysis or evaluation of the issues and events is either vague or *completely* inaccurate.
Clarity (25%)	A non-expert (e.g., a fellow student) can understand the solutions. All concepts and used techniques are defined and explained. Whenever it is applicable, the solution is accompanied by illustrative plots that are explained and interpreted. Accompanied code is well commented and easy to follow.	The teacher (or other professional physicists) can understand the solution but a non-expert might have some trouble doing so. The solution has some minor shortcuts or some non-explained assumptions. Not every step of the analysis is explained, but it is still possible to follow the author's logic. The code is not well commented but it is still possible to follow it.	It is hard to follow the solutions. The solution has some major shortcuts and hidden assumptions. The analysis or evaluation of the issues and events is vague. The code is not well commented but it is either hard or impossible to follow it.
Validity (25%)	All calculations are correct. The final values are right.	Small mistake in the code and/or calculations (e.g., a	Major mistakes in the code and/or in the analysis. The

	The interpretations and final concussions are valid.	wrong sign, a missing constant). The final answer is close to the correct value (e.g., by a small factor; twice too large or twice too small; however, the general trend is correct).	final values and conclusions are incorrect.

Course Project

An integral part of this course is the course project, which builds on the topics and techniques covered in the class, focusing on extending and evaluating methods to solve problems. Students will write a written proposal for the project, carry out literature review, conduct the project, review two proposals by your peers, address the comments you received on your own peer-reviews, and then write a paper about the project, and present the project in class. Students are encouraged to identify a new problem, apply or extend the methods they learned in class to propose an approach to solve the problem. Students will propose a novel project, do the research and build a proof-of-concept, write a report about the work, and present the work in class. Emphasis is placed on quantitative evaluation of the approach. Each student will team up with a classmate (max of 2 allowed in some cases) to do an independent project based on the topics covered in the class.

The objective of this assignment is to a) explore literature regarding data science and machine learning, b) synthesize the acquired knowledge in the form of article, c) learn how to write peer-review comments, d) learn how to respond to peer-review comments, e) be able to summarize a weeks-long project in a form of a condensed, short presentation.

Structure and Formatting:

We encourage you to use the LaTeX template https://www.overleaf.com/read/crjbtrfftfhg that we prepared for you in Overleaf. If you use a WYSIWYG editor, please remember to submit your article in the PDF format (not docx, rtf or odt). In the paper, you must provide a link to a GitHub repository with the relevant code, scripts, or notebooks. Python 3.8+ is preferred, but in principle, you are free to use any language of your choice - as long as the code is clear and well commented (the reader should be able to go to your repository, clone your repository and run your code without getting any errors).

Project Timeline:

- Week 1-2: Identify team members.
- Week 3-4: Choose your topic. Prepare and post a work plan/proposal **by Wednesday, February 2, at 4 p.m**.
- Week 6: Find relevant literature. Read about your topic. Prepare a literature review **by Wednesday, February 16, at 4 p.m**.
- Week 6-8: Make a plan for your article. Decide which aspects you are going to describe and which leave out. After all, you have limited space (only a couple of pages, including figures and bibliography). Complete the necessary coding and calculations. Prepare plots and figures.
- Week 9: Write the first version of your report. You should have a draft by March 9. Proofread your report. Make sure that all key terms are defined. Make sure that the report has the right structure (abstract, introduction, the main content, discussion/summary, and bibliography). Remember, that the list of references at the end of your report is not enough - your sources must be cited in the article.

- Week 10: Prepare a pdf of your article (this is your Midterm Report). Make sure that the number of words is below the maximum limit (3,500). Make sure that your name, affiliation, abstract and paper title are visible on the first page. Submit the pdf using a Blackboard **by Wednesday, March 23, not later than 4 p.m.**
- Week 12: Choose two reports prepared by your peers. Read those reports. Using Piazza, give each author suggestions on how they can improve the papers. To make sure that each person will receive an equal number of comments, only the first two comments under each project will count for credit (though you are still welcome to give comments to more than two papers if you wish; it will just not count as extra credit). You should complete this action **by Wednesday, April 6, at 4 p.m.**
- Read the suggestions that you received from your peers. Address them (either incorporate the suggested changes or challenge them, describing why you think those changes would not improve the quality of your article).
- Prepare an in-class presentation (length depends on the number of projects, but plan 5-7 min presentation) **on April 25 and 27th (in class).**
- Submit your final report **by Wednesday, May 4 at 4:00 p.m.**

Sample project:

"Ageism in Traffic Policing:" the goal of the project is to explore the nature of police stops in the USA, with a particular focus on bias. Using a large corpus of police stops, the students conducted an analysis of the features that are indicative of the outcome of a traffic stop. They found that ageism is rampant in traffic policing with older subjects more likely to receive lenient outcomes. Students found that some patterns are correlated with the political leaning of the state, with "red" states presenting less ageism than "blue" states.

Grading Breakdown

Homework: There will be 4 homework assignments.
Quizzes: While there are 5 quizzes, only the 4 best scores are used for grade calculation.
Class Participation: Students are expected to attend every class and actively participate in the discussion.
Project: Projects will be graded on novelty, technical soundness, and the quality of evaluation. Reports and presentations will be graded according to the project grading rubric and the quality and clarity of presentation.

Assignment	Points	% of Grade
Homeworks	40	40
Quizzes	20	20
Class participation	5	5
Project proposal	1	1
Literature review	2	2
Midterm report	5	5
Peer review	2	2
Project report	15	15

Project presentation	10	10
TOTAL	100	100

Assignment Submission Policy

Homework assignments are due at 4pm on the due date and should be submitted on Blackboard. A total of 7 late days can be used. For each submission, a maximum of 2 late days are allowed. After that each day you will lose 20% of the possible points for the assignment. After one week, the assignment cannot be submitted. No late days are allowed for quizzes. Late submission of quizzes will receive 0 points.

Course Schedule: A Weekly Breakdown (next page)

	Topics/Daily Activities	Readings and Homework *It is expected that students read these in advance of each week and come to class prepared to discuss them.*	Deliverable / Due Dates
Week 1 Jan 10, 12 Prof. Lerman	**Course Introduction & Research Ethics**	● Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*. ● Olteanu, Castillo & Kiciman "Social data: Biases, methodological pitfalls, and ethical boundaries." ● Angwin et al. Machine bias. Propublica (2016) https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing	Homework 1 Assigned
Week 2 Jan 19 Prof. Morstatter	**Unsupervised methods: clustering, PCA, etc**	● Grira, Nizar, Michel Crucianu, and Nozha Boujemaa. "Unsupervised and semi-supervised clustering: a brief survey." A review of machine learning techniques for processing multimedia content 1 (2004): 9-16.	Homework 1 Due
Week 3 Jan 24, 26 Prof. Lerman	**Supervised methods: Regression, Explanation, etc.**	● Bodo Winter, Tutorial on linear models http://www.bodowinter.com/tutorial/bw_LME_tutorial1.pdf ● Bodo Winter, Tutorial on mixed effects analysis, http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf ● Shapley ● Fennell, Zou & Lerman (2019) "Explaining and Predicting Behavioral Data with Structured Feature Space Decomposition", in *EPJ Data Science*	Homework 2 Assigned
Week 4 Jan 31, Feb 2 Prof. Morstatter	**Definitions and fair machine learning**	● Kleinberg, Mullainathan, and Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores (2016) https://arxiv.org/abs/1609.05807 ● Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In Advances in Neural Information Processing Systems (pp. 5680-5689). http://papers.nips.cc/paper/7151-on-fairness-and-calibration ● Corbett-Davies, Sam, et al. "Algorithmic decision making and the cost of fairness." Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.	Project Proposals Due

		ACM, 2017. https://dl.acm.org/citation.cfm?id=3098095	
Week 5 Feb 7, Prof. Morstatter Feb 9 Prof. Lerman	**Bias in Data**	● Ruths & Pfeffer "Social Media for Large Studies of Behavior" ● Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671. Part 1. http://www.cs.yale.edu/homes/jf/BarocasSelbst.pdf ● Vaupel, J. W., & Yashin, A. I. (1985). Heterogeneity's ruses: some surprising effects of selection on population dynamics. *The American Statistician*, *39*(3), 176-185. https://www.jstor.org/stable/2683925?seq=1#metadata_info_tab_contents ● Lerman, K. (2018). Computational social scientist beware: Simpson's paradox in behavioral data. *Journal of Computational Social Science*, *1*(1), 49-58. https://arxiv.org/abs/1710.08615 ● Alipourfard, N., Fennell, P. G., & Lerman, K. (2018). Using Simpson's Paradox to Discover Interesting Patterns in Behavioral Data. In *Twelfth International AAAI Conference on Web and Social Media*. https://link.springer.com/article/10.1007/s42001-017-0007-4	Homework 2 due Homework 3 assigned
Week 6 Feb 14, 16 Prof. Morstatter	**Basics of NLP: Text mining and word embeddings**	● Bolukbasi, Chang, Zou, Saligrama & Kalai "Man is to computer programmer as woman is to homemaker? debiasing word embeddings" ● Zhao, Wang, Yatskar, Ordonez & Chang "Men also like shopping: Reducing gender bias amplification using corpus-level constraints" ● Zhao, Wang, Yatskar, Ordonez & Chan "Gender bias in coreference resolution: Evaluation and debiasing methods"	Lit review due
Week 7 Feb 23 Prof. Lerman Option: Nina guest lecture	**Bias and Fairness in NLP**	● Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326). https://lexitron.nectec.or.th/public/LREC-2010_Malta/pdf/385_Paper.pdf ● Arvind Narayanan	Homework 3 due Homework 4 assigned

		● S. Golder and M. Macy, Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures, Science Vol. 333 no. 6051 pp. 1878-1881, 2011. https://science.sciencemag.org/content/333/6051/1878/ ● **More... Language reveals** ● Mehrabi, Ninareh, et al. "Man is to person as woman is to location: Measuring gender bias in named entity recognition." *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. 2020. ● Zhao, Jieyu, et al. "Men also like shopping: Reducing gender bias amplification using corpus-level constraints." arXiv preprint arXiv:1707.09457 (2017).	
Week 8 Feb 28, Mar 2 Prof. Morstatter	**Basics of Networks**	● Easley & Kleinberg "Networks, crowds, and markets" [Chapter 2] ● Fortunato & Hric "Community detection in networks: A user guide" ● Mehrabi, Morstatter, Peng & Galstyan "Debiasing Community Detection: The Importance of Lowly-Connected Nodes"	
Week 9 Mar 7, 9 Prof. Lerman	**Fairness & Bias in Networks**	● Kooti, F., Hodas, N. O., & Lerman, K. (2014). Network weirdness: Exploring the origins of network paradoxes. In *Eighth International AAAI Conference on Weblogs and Social Media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewFile/8101/8127 ● Lee, E., Karimi, F., Wagner, C., Jo, H. H., Strohmaier, M., & Galesic, M. (2019). Homophily and minority-group size explain perception biases in social networks. *Nature human behaviour*, 1-10. https://www.nature.com/articles/s41562-019-0677-4 ● D Liben-Nowell & J Kleinberg, "The link prediction problem for social networks." Journal of the American Society for Information Science and Technology, Vol. 58, No. 7. (May 2007), pp. 1019-1031. ● Stoica, AA, Riederer, C and Chaintreau, A. 2018. Algorithmic Glass Ceiling in Social Networks: The effects of social	Homework 4 due

		recommendations on network diversity. In Proceedings of the 2018 World Wide Web Conference (WWW '18). 923-932. DOI: https://doi.org/10.1145/3178876.3186140	
Mar 14, 16		**SPRING BREAK**	
Week 10 Mar 21, 23 Prof. Lerman	**Debiasing Methods**	● Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." Knowledge and Information Systems 33.1 (2012): 1-33. ● Zhang et al. "Mitigating unwanted biases with adversarial learning." Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018.	Midterm report due
Week 11 Mar 28 Prof. Morstatter Mar 30, TA	**Fair Representations**	● Adversarial approaches ● Fair representation of graphs DeepWalk, FairWalk, ● Fair representation of NLP, etc. BERT, Fair Word embedding	
Week 12 Apr 4, 6 Prof. Morstatter	**Privacy & Fraud**	● Ferrara, Varol, Davis, Menczer & Flammini "The rise of social bots". ● Pfeffer, Mayer & Morstatter "Tampering with Twitter's Sample API". ● Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences, 110(15):5802-5805 ● Zheleva E., Getoor L. (2011) Privacy in Social Networks: A Survey. In: Aggarwal C. (eds) Social Network Data Analytics. Springer, Boston, MA https://link.springer.com/chapter/10.1007/978-1-4419-8462-3_10	Peer reviews due
Week 13 Apr 11, 13 Prof. Lerman	**Algorithmic Bias and Feedback**	● Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *science*, *311*(5762), 854-856. https://science.sciencemag.org/content/311/5762/854 ● Ricardo Baeza-Yates "Bias on the Web" Communications of the ACM, June 2018, Vol. 61 No. 6, Pages 54-61 https://www.researchgate.net/profile/Ricardo_Baeza-Yates/publication/325330277_Bias_on_the_web/links/5b1576440f7e9bda0ffcc999/Bias-on-the-web.pdf	

		● Lerman, K., & Hogg, T. (2014). Leveraging position bias to improve peer recommendation. *PloS one*, *9*(6), e98914. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098914	
Week 14 Apr 18, 20 Prof. Lerman	**Researcher Bias & Bonus Topic**	● MacCoun, R., & Perlmutter, S. (2015). Blind analysis: hide results to seek the truth. *Nature News*, *526*(7572), 187. https://www.nature.com/news/blind-analysis-hide-results-to-seek-the-truth-1.18510 ● Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E. C., … Nosek, B. A. (2017, April 24). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. https://doi.org/10.1177/2515245917747646 ● Arvind Narayanan's talk "How to recognize AI snake oil" https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf	
Week 15 Apr 25, 27	**Project Presentations**		Project Presentation slides due
FINAL May 4	**Project Final Report**		Project Final Report due

Statement on Academic Conduct and Support Systems

Academic Conduct:

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in SCampus in Part B, Section 11, "Behavior Violating University Standards" policy.usc.edu/scampus-part-b. Other forms of academic dishonesty are equally unacceptable. See additional information in SCampus and university policies on scientific misconduct, policy.usc.edu/scientific-misconduct.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
studenthealth.usc.edu/counseling
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

National Suicide Prevention Lifeline - 1 (800) 273-8255 – 24/7 on call
suicidepreventionlifeline.org
Free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week.

Relationship and Sexual Violence Prevention and Services (RSVP) - (213) 740-9355(WELL), press "0" after hours – 24/7 on call
studenthealth.usc.edu/sexual-assault
Free and confidential therapy services, workshops, and training for situations related to gender-based harm.

Office of Equity and Diversity (OED)- (213) 740-5086 | Title IX – (213) 821-8298
equity.usc.edu, titleix.usc.edu
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants. The university prohibits discrimination or harassment based on the following *protected characteristics*: race, color, national origin, ancestry, religion, sex, gender, gender identity, gender expression, sexual orientation, age, physical disability, medical condition, mental disability, marital status, pregnancy, veteran status, genetic information, and any other characteristic which may be specified in applicable laws and governmental regulations. The university also prohibits sexual assault, non-consensual sexual contact, sexual misconduct, intimate partner violence, stalking, malicious dissuasion, retaliation, and violation of interim measures.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
usc-advocate.symplicity.com/care_report
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office of Equity and Diversity |Title IX for appropriate investigation, supportive measures, and response.

The Office of Disability Services and Programs - (213) 740-0776
dsp.usc.edu
Support and accommodations for students with disabilities. Services include assistance in providing readers/notetakers/interpreters, special accommodations for test taking needs, assistance with architectural barriers, assistive technology, and support for individual needs.

USC Support and Advocacy - (213) 821-4710
uscsa.usc.edu
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity at USC - (213) 740-2101
diversity.usc.edu
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
dps.usc.edu, emergency.usc.edu
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-120 – 24/7 on call
dps.usc.edu
Non-emergency assistance or information.