Fairness and Bias in Peer Review and other Sociotechnical Intelligent Systems (AAAI 2020 Tutorial Syllabus)

Nihar B. Shah and Zachary Lipton
Carnegie Mellon University

The tutorial is organized into two parts. In the **first part**, Zachary Lipton will articulate current and historical thinking on the social impacts of applied machine learning. Calling upon the economics literature on statistical discrimination and the more recent literature on fairness in machine learning, he will present a critical survey of attempts by academics to formally analyze and mitigate these problems. Throughout, technical formulations will be presented alongside real-world motivations. Technical formulations and solutions will be accompanied by critical discussion, calling attention to gaps between legal doctrine, ethical principles, and the reductive technical definitions intended to capture them, highlighting the ways that purported technical fixes may themselves have the potential to confer harm, e.g., by missing the point entirely, by obfuscating the critical questions, by codifying problematic concept (e.g., race), and by misleading policy makers with apparent solutions that do not actually solve the policy problems that they purport to address.

Outline of part 1 of the tutorial including references:

1. *Historical context:* We will discuss conceptions of bias and fairness broadly as construed in ethical and legal frameworks. Will address both procedural fairness and notions of fairness concerning group membership.
2. *Economic frameworks:* Next, we will introduce the classic literature on fairness in hiring due to economists, including the Becker and Phelps models of taste-based and statistical discrimination respectively (Bec57; Phe72; AC77; A$^+$73). We will also cover recent extensions from the ML community to classic economic models (HC18; CLM19).
3. *Automated decisions:* To set up a discussion of the ML fairness work, we will motivate modern issues related to predictive technology as used in lending, resume screening, recommender systems, and recidivism prediction systems used in criminal justice.
4. *Fair machine learning:* Next, we will discuss attempts by the machine learning community to formalize notions of fairness in the context of classification. We will describe various parity measures that have served as "definitions of fairness" in rigorous mathematical study, covering both associative and counterfactual measures (HPS$^+$16; DHP$^+$12; ZWS$^+$13; Cho17; KR18; LMC18; KLRS17; KCP$^+$17; NS18).
5. *Limitations:* The first part of the tutorial will conclude with a critical discussion of work to date, highlighting the gaps left between underlying social desiderata and reductive technical definitions. The discussion will also highlight some of the perils of a form of overclaiming that tends to slip past ML peer review: representing to have made substantial progress on a pressing social problem without in fact offering a viable solution.

In the **second part**, Nihar Shah will discuss biases due to factors such as subjectivity, calibration, strategic behavior in human-provided data. Applications in focus here include peer review, hiring, admissions, peer grading, A/B testing, crowdsourcing, and online ratings.

Specifically, this part will use peer review as a running example application because: (i) We envisage that most members of the audience at AAAI would be cognizant of peer review, and a large fraction would have had a first hand experience. (ii) To the best of our knowledge, no tutorial in ML/AI in the last several years focuses on this application.

Outline of part 2 of the tutorial including references:

1. *Biases:* We will make a smooth transition into peer review from part 1, by first discussing biases due to demographics in single-blind peer review. We will discuss a remarkable randomized controlled trial (TZH17) at the WSDM 2017 conference, and associated hypothesis testing problems. Auxiliary references: (OHKL16; BTA$^+$08; WOF08; HJP03; SSS19).
2. *Noise:* By noise, here we mean poor reviews due to inappropriate choice of reviewers. We will overview widely used reviewer assignment algorithms (CZ13), its shortcomings, and recent research focusing on fairness (SSS18; KSM19). Auxilliary references: (RBVdS07; MM07; LSM14; RB08; TCH17; GS07; TTT10; CZB12; LWPY13; GKK$^+$10; BL01; HWC99; FSR19).
3. *Subjectivity:* Unfairness due to subjective opinions of individual evaluators, and using ML + social choice theory to mitigate it (KTP77; NSP18; Lee15).Will discuss fundamental theory and empirical evaluation on IJCAI 2017.
4. *Miscalibration:* Unfairness due to miscalibrations (e.g.,

strictness, leniency, extremal behavior) of the evaluator (GWG13; WS19), and using ML+information theory to mitigate it. Auxilliary references: (Pau81; BK13; GWG13; MKLP17; Pau81; RRS11; SBGW17; Sha17; FSG+10).

5. *Strategic behavior:* Unfairness if some entities gain advantage by gaming the system in a zero sum game setting like in peer review, college admissions, and hiring. We will present an experiment from (BGH16) and overview an algorithmic building block that is common to (AFPT11; DCMT08; HM13; FK15; KLMP15; ALM+16; KKK+17; XZSS18).

6. *Policy:* The presentation will conclude with a discussion on driving actual policy change.

The presentation will be interspersed with empirical analyses of NeurIPS 2016 peer review (STM+17).

References

[A+73] Kenneth Arrow et al. The theory of discrimination. *Discrimination in labor markets*, 3(10):3–33, 1973.

[AC77] Dennis J Aigner and Glen G Cain. Statistical theories of discrimination in labor markets. *ILR Review*, 30(2):175–187, 1977.

[AFPT11] Noga Alon, Felix Fischer, Ariel Procaccia, and Moshe Tennenholtz. Sum of us: Strategyproof selection from the selectors. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 101–110. ACM, 2011.

[ALM+16] Haris Aziz, Omer Lev, Nicholas Mattei, Jeffrey S Rosenschein, and Toby Walsh. Strategyproof peer selection: Mechanisms, analyses, and experiments. In *AAAI*, pages 397–403, 2016.

[Bec57] Gary S Becker. The economics of discrimination chicago. *University of Chicago*, 1957.

[BGH16] Stefano Balietti, Robert L Goldstone, and Dirk Helbing. Peer review and competition in the art exhibition game. *Proceedings of the National Academy of Sciences*, 113(30):8414–8419, 2016.

[BK13] Yukino Baba and Hisashi Kashima. Statistical quality estimation for general crowdsourcing tasks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013.

[BL01] Salem Benferhat and Jerome Lang. Conference paper assignment. *International Journal of Intelligent Systems*, 16(10):1183–1192, 10 2001.

[BTA+08] Amber E. Budden, Tom Tregenza, Lonnie W. Aarssen, Julia Koricheva, Roosa Leimu, and Christopher J. Lortie. Double-blind review favours increased representation of female authors. *Trends in Ecology and Evolution*, 23(1):4 – 6, 2008.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[CLM19] Lee Cohen, Zachary C Lipton, and Yishay Mansour. Efficient candidate screening under multiple tests and implications for fairness. *arXiv preprint arXiv:1905.11361*, 2019.

[CZ13] L. Charlin and R. S. Zemel. The Toronto Paper Matching System: An automated paper-reviewer assignment system. In *ICML Workshop on Peer Reviewing and Publishing Models*, 2013.

[CZB12] L. Charlin, R. S. Zemel, and C. Boutilier. A framework for optimizing paper matching. *CoRR*, abs/1202.3706, 2012.

[DCMT08] Geoffroy De Clippel, Herve Moulin, and Nicolaus Tideman. Impartial division of a dollar. *Journal of Economic Theory*, 139(1):176–191, 2008.

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, 2012.

[FK15] Felix Fischer and Max Klimm. Optimal impartial selection. *SIAM Journal on Computing*, 44(5):1263–1285, 2015.

[FSG+10] Peter A. Flach, Sebastian Spiegler, Bruno Golénia, Simon Price, John Guiver, Ralf Herbrich, Thore Graepel, and Mohammed J. Zaki. Novel tools to streamline the conference review process: Experiences from SIGKDD'09. *SIGKDD Explor. Newsl.*, 11(2):63–67, May 2010.

[FSR19] T Fiez, N Shah, and L Ratliff. A SUPER* algorithm to optimize paper bidding in peer review. In *ICML workshop on Real-world Sequential Decision Making: Reinforcement Learning And Beyond*, 2019.

[GKK+10] N. Garg, T. Kavitha, A. Kumar, K. Mehlhorn, and J. Mestre. Assigning papers to referees. *Algorithmica*, 58(1):119–136, Sep 2010.

[GS07] Judy Goldsmith and Robert H. Sloan. The AI conference paper assignment problem. WS-07-10:53–57, 12 2007.

[GWG13] Hong Ge, Max Welling, and Zoubin Ghahramani. A Bayesian model for calibrating conference review scores, 2013.

[HC18] Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1389–1398. International World Wide Web Conferences Steering Committee, 2018.

[HJP03] Shawndra Hill and Foster J. Provost. The myth of the double-blind review? author identification using only citations. *SIGKDD Explorations*, 5:179–184, 01 2003.

[HM13] Ron Holzman and Hervé Moulin. Impartial nominations for a prize. *Econometrica*, 81(1):173–196, 2013.

[HPS+16] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[HWC99] David Hartvigsen, Jerry C. Wei, and Richard Czuchlewski. The conference paper-reviewer assignment problem. *Decision Sciences*, 30(3):865–876, 1999.

[KCP+17] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[KKK+17] Anson B Kahng, Yasmine Kotturi, Chinmay Kulkarni, David Kurokawa, and Ariel D. Procaccia. Ranking wily people who rank each other. *Technical Report*, 2017.

[KLMP15] David Kurokawa, Omer Lev, Jamie Morgenstern, and Ariel D Procaccia. Impartial peer review. In *IJCAI*, pages 582–588, 2015.

[KLRS17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[KR18] Jon Kleinberg and Manish Raghavan. Selection problems in the presence of implicit bias. *arXiv preprint arXiv:1801.03533*, 2018.

[KSM19] Ari Kobren, Barna Saha, and Andrew McCallum. Paper matching with local fairness constraints. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

[KTP77] Steven Kerr, James Tolliver, and Doretta Petree. Manuscript characteristics which influence acceptance for management and social science journals. *Academy of Management Journal*, 20(1):132–141, 1977.

[Lee15] Carole J Lee. Commensuration bias in peer review. *Philosophy of Science*, 82(5):1272–1283, 2015.

[LMC18] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pages 8125–8135, 2018.

[LSM14] Xiang Liu, Torsten Suel, and Nasir Memon. A robust model for paper reviewer assignment. In *ACM Conference on Recommender Systems*, 2014.

[LWPY13] Cheng Long, Raymond Wong, Yu Peng, and Liangliang Ye. On good and fair paper-reviewer assignment. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 1145–1150, 12 2013.

[MKLP17] R. S. MacKay, R. Kenna, R. J. Low, and S. Parker. Calibration with confidence: a principled method for panel assessment. *Royal Society Open Science*, 4(2), 2017.

[MM07] David Mimno and Andrew McCallum. Expertise modeling for matching papers with reviewers. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007.

[NS18] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access, 2018.

[NSP18] Ritesh Noothigattu, Nihar Shah, and Ariel Procaccia. Choosing how to choose papers. *arXiv preprint arxiv:1808.09057*, 2018.

[OHKL16] Kanu Okike, Kevin T. Hug, Mininder S. Kocher, and Seth S. Leopold. Single-blind vs Double-blind Peer Review in the Setting of Author PrestigeSingle-blind vs Double-blind Peer Review in the Setting of Author PrestigeLetters. *JAMA*, 316(12):1315–1316, 09 2016.

[Pau81] S. R. Paul. Bayesian methods for calibration of examiners. *British Journal of Mathematical and Statistical Psychology*, 34(2):213–223, 1981.

[Phe72] Edmund S Phelps. The statistical theory of racism and sexism. *The american economic review*, pages 659–661, 1972.

[RB08] Marko A. Rodriguez and Johan Bollen. An algorithm to determine peer-reviewers. In *ACM Conference on Information and Knowledge Management*, 2008.

[RBVdS07] Marko A Rodriguez, Johan Bollen, and Herbert Van de Sompel. Mapping the bid behavior of conference referees. *Journal of Informetrics*, 1(1):68–82, 2007.

[RRS11] Magnus Roos, Jörg Rothe, and Björn Scheuermann. How to calibrate the scores of biased reviewers by quadratic programming. In *AAAI Conference on Artificial Intelligence*, 2011.

[SBGW17] Nihar B. Shah, Sivaraman Balakrishnan, Adityanand Guntuboyina, and Martin J. Wainwright. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Transactions on Information Theory*, 63(2):934–959, 2017.

[Sha17] Nihar Shah. *Learning From People*. PhD thesis, PhD thesis, EECS Department, University of California, Berkeley, 2017.

[SSS18] Ivan Stelmakh, Nihar Shah, and Aarti Singh. PeerReview4All: Fair and accurate reviewer assignment in peer review. *arXiv preprint arxiv:1806.06237*, 2018.

[SSS19] Ivan Stelmakh, Nihar Shah, and Aarti Singh. On testing for biases in peer review. In *ACM EC workshop on Mechanism Design for Social Good*, 2019.

[STM+17] Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. Design and analysis of the nips 2016 review process. *arXiv preprint arXiv:1708.09794*, 2017.

[TCH17] H. D. Tran, G. Cabanac, and G. Hubert. Expert suggestion for conference program committees. In *2017 11th International Conference on Research Challenges in Information Science (RCIS)*, pages 221–232, May 2017.

[TTT10] Wenbin Tang, Jie Tang, and Chenhao Tan. Expertise matching via constraint-based optimization. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010.

[TZH17] Andrew Tomkins, Min Zhang, and William D. Heavlin. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708–12713, 2017.

[WOF08] Thomas J. Webb, Bob OHara, and Robert P. Freckleton. Does double-blind review benefit female authors? *Trends in Ecology and Evolution*, 23(7):351 – 353, 2008.

[WS19] Jingyan Wang and Nihar B Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *AAMAS*, 2019.

[XZSS18] Yichong Xu, Han Zhao, Xiaofei Shi, and Nihar Shah. On strategyproof conference review. *arXiv preprint arxiv:1806.06266*, 2018.

[ZWS+13] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, 2013.