TheWebConf 2020 Tutorial: Fairness and Bias in Peer Review and other Sociotechnical Intelligent Systems

Nihar B. Shah
nihars@cs.cmu.edu
Carnegie Mellon University

Zachary Lipton
zlipton@cs.cmu.edu
Carnegie Mellon University

ABSTRACT

Questions of fairness and bias abound in all socially-consequential decision-making. Whether designing the protocols for peer review of research papers, determining which job candidates surface in a search, or deciding who should receive a loan in an online lending platform, any decision with the potential to allocate benefits or confer harms raises concerns about *who* gains or loses that may fail to surface in naively-chosen performance measures. Data science interacts with these questions in two fundamentally different ways: (i) as the technology driving the very systems responsible for certain social impacts, posing new questions about what it means for such systems to accord with ethical norms and the law; and (ii) as a set of powerful tools for analyzing existing systems (even those that don't themselves depend on ML), e.g. for auditing existing systems for various biases.

This tutorial will tackle both angles on the interaction between technology and society vis-a-vis concerns over fairness and bias. Our presentation will cover a wide range of disciplinary perspectives. The first part will focus on the social impacts of technology and the formulations of fairness and bias defined via protected characteristics. The second part will focus on peer review and other web-related applications, and explore other forms of bias, such as that due to subjectivity, miscalibration, and fraudulent behavior.

PART I OF THE TUTORIAL

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, focusing on decision-making in various online platforms, including news feeds, job-finding sites, and lending. 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 and critical discussion, calling attention to the gaps between legal doctrine, ethical principles, and the technical definitions intended to capture them. This section will also highlight some ways that purported fixes can themselves confer harm, e.g., by obfuscating the critical questions, by codifying problematic concepts (e.g., race), and by creating incentive structures that exacerbate the problems that they were intended to mitigate.

Outline of part 1 of the tutorial:

(1) *Historical context:* We will discuss conceptions of bias and fairness through dominant ethical and legal frameworks. This discussion will contextualize various concepts in US Civil Rights law that are subject to frequent but glancing references in recent papers on fair ML.
(2) *Economic frameworks:* 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 [1, 3, 7, 40]. We will also cover recent extensions from the ML community to classic economic models [12, 24].
(3) *Automated decisions:* To motivate the discussion fairness in ML systems, we discuss its application in various online services, including lending platforms, recommender systems, news feeds, job-finding platforms, etc.
(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 been presented as "definitions of fairness" in rigorous mathematical study, covering both associative and counterfactual measures [11, 14, 21, 27, 28, 31, 33, 36, 55].
(5) The first part of the tutorial will conclude with a critical discussion of work to date. The discussion will highlight the perils of solutionism, where papers representing to have made substantial progress on pressing social problems (when in fact they have not) are picked up by companies and represented as certificates of fairness.

PART II OF THE TUTORIAL

In the second part, Nihar Shah will discuss issues of bias and unfairness due to factors such as subjectivity, calibration, strategic behavior in human-provided data. Applications in focus here include recommendation systems, crowdsourcing, online rating systems, A/B testing over the Internet, peer grading, and hiring. This part will use the application of peer review as a running example application. We envisage that most TheWebConf2020 attendees will be cognizant of peer review and that a large fraction would have first-hand experience with the process.

Outline of part 2 of the tutorial:

(1) *Demographics:* We will make a smooth transition from part 1 into peer review, by first discussing biases due to demographics in single-blind peer review. We will discuss a remarkable randomized controlled trial [49] at the WSDM 2017 conference, and associated hypothesis testing problems [46]. Auxiliary references: [8, 22, 38, 53].
(2) *Assignment of reviewers:* We will detail the current methods of assigning reviewers to papers in major ML/AI conferences [9]. We will then highlight problems of unfairness therein, and discuss alternative assignments with theoretical guarantees [18, 29, 47], and empirical evaluations on CVPR 2017, CVPR 2018, and MIDL 2018 [29]. Auxilliary references: [10, 15, 20, 34, 48].
(3) *Subjectivity:* Unfairness due to subjective opinions of individual evaluators, and using ML + social choice theory to mitigate

it [26, 32, 37]. We will discuss fundamental theory as well as empirical evaluation on IJCAI 2017.

(4) *Miscalibration:* Unfairness due to miscalibrations (e.g., strictness, leniency, extremal behavior) of the evaluator [19, 52], and using ML+information theory to mitigate it. Auxilliary references: [5, 17, 19, 35, 39, 39, 41–43, 45].
(5) *Fraudulent 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 [6] and discuss an algorithmic building block that is common to many works [2, 4, 13, 16, 23, 25, 30, 54].
(6) *Policy:* The presentation will conclude with a discussion on driving actual policy change [50, 51].

The presentation will be interspersed with empirical analyses of NeurIPS 2016 peer review [44].

REFERENCES

[1] Dennis J Aigner and Glen G Cain. 1977. Statistical theories of discrimination in labor markets. *ILR Review* 30, 2 (1977), 175–187.
[2] Noga Alon, Felix Fischer, Ariel Procaccia, and Moshe Tennenholtz. 2011. Sum of us: Strategyproof selection from the selectors. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*. ACM, 101–110.
[3] Kenneth Arrow et al. 1973. The theory of discrimination. *Discrimination in labor markets* 3, 10 (1973), 3–33.
[4] H Aziz, O Lev, N Mattei, J Rosenschein, and T Walsh. 2016. Strategyproof Peer Selection: Mechanisms, Analyses, and Experiments.. In *AAAI*. 397–403.
[5] Yukino Baba and Hisashi Kashima. 2013. Statistical Quality Estimation for General Crowdsourcing Tasks. In *KDD*.
[6] S Balietti, R Goldstone, and D Helbing. 2016. Peer review and competition in the Art Exhibition Game. *Proceedings of the National Academy of Sciences* (2016).
[7] G Becker. 1957. The economics of discrimination Chicago. *U. Chicago* (1957).
[8] A Budden, T Tregenza, L Aarssen, J Koricheva, R Leimu, and C Lortie. 2008. Double-blind review favours increased representation of female authors. *Trends in Ecology and Evolution* (2008).
[9] L. Charlin and R. S. Zemel. 2013. The Toronto Paper Matching System: An automated paper-reviewer assignment system. In *ICML Workshop on Peer Reviewing and Publishing Models*.
[10] L. Charlin, R. S. Zemel, and C. Boutilier. 2012. A Framework for Optimizing Paper Matching. *CoRR* abs/1202.3706 (2012). http://arxiv.org/abs/1202.3706
[11] A Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* (2017).
[12] L Cohen, Z Lipton, and Y Mansour. 2019. Efficient candidate screening under multiple tests and implications for fairness. *arXiv:1905.11361* (2019).
[13] Geoffroy De Clippel, Herve Moulin, and Nicolaus Tideman. 2008. Impartial division of a dollar. *Journal of Economic Theory* 139, 1 (2008), 176–191.
[14] C Dwork, M Hardt, T Pitassi, O Reingold, and R Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*.
[15] T Fiez, N Shah, and L Ratliff. 2019. A SUPER* Algorithm to Optimize Paper Bidding in Peer Review. In *ICML workshop on Real-world Sequential Decision Making: Reinforcement Learning And Beyond*.
[16] F Fischer and M Klimm. 2015. Optimal impartial selection. *SIAM J. Comput.* (2015).
[17] P Flach, S Spiegler, B Golénia, S Price, J Guiver, R Herbrich, T Graepel, and M Zaki. 2010. Novel Tools to Streamline the Conference Review Process: Experiences from SIGKDD'09. *SIGKDD Explor. Newsl.* (2010).
[18] N. Garg, T. Kavitha, A. Kumar, K. Mehlhorn, and J. Mestre. 2010. Assigning Papers to Referees. *Algorithmica* 58, 1 (01 Sep 2010), 119–136.
[19] H Ge, M Welling, and Z Ghahramani. 2013. A Bayesian model for calibrating conference review scores. http://mlg.eng.cam.ac.uk/hong/nipsrevcal.pdf
[20] Judy Goldsmith and Robert H. Sloan. 2007. The AI conference paper assignment problem. WS-07-10 (12 2007), 53–57.
[21] M Hardt, E Price, and N Srebro. 2016. Equality of opportunity in supervised learning. In *NeurIPS*.
[22] S Hill and F Provost. 2003. The myth of the double-blind review? Author identification using only citations. *SIGKDD Explorations* (2003).
[23] R Holzman and H Moulin. 2013. Impartial nominations for a prize. *Econometrica* (2013).
[24] L Hu and Y Chen. 2018. A short-term intervention for long-term fairness in the labor market. In *WWW*.
[25] A Kahng, Y Kotturi, C Kulkarni, D Kurokawa, and A Procaccia. 2017. Ranking Wily People Who Rank Each Other. *Technical Report* (2017).
[26] S Kerr, J Tolliver, and D Petree. 1977. Manuscript characteristics which influence acceptance for management and social science journals. *Acad. Mgmt. Jrnl.* (1977).
[27] N Kilbertus, M Carulla, G Parascandolo, M Hardt, D Janzing, and B Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *NeurIPS*.
[28] Jon Kleinberg and Manish Raghavan. 2018. Selection Problems in the Presence of Implicit Bias. *arXiv preprint arXiv:1801.03533* (2018).
[29] A Kobren, B Saha, and A McCallum. 2019. Paper Matching with Local Fairness Constraints. In *KDD*.
[30] D Kurokawa, O Lev, J Morgenstern, and A Procaccia. 2015. Impartial Peer Review. In *IJCAI*.
[31] M Kusner, J Loftus, C Russell, and R Silva. 2017. Counterfactual fairness. In *NeurIPS*. 4066–4076.
[32] C Lee. 2015. Commensuration bias in peer review. *Philosophy of Science* (2015).
[33] Z Lipton, J McAuley, and A Chouldechova. 2018. Does mitigating ML's impact disparity require treatment disparity?. In *NeurIPS*.
[34] Cheng Long, R Wong, Y Peng, and L Ye. 2013. On Good and Fair Paper-Reviewer Assignment.
[35] R. MacKay, R. Kenna, R. Low, and S. Parker. 2017. Calibration with confidence: a principled method for panel assessment. *Royal Society Open Science* (2017).
[36] R Nabi and I Shpitser. 2018. Fair inference on outcomes. In *AAAI*.
[37] Ritesh Noothigattu, Nihar Shah, and Ariel Procaccia. 2018. Choosing how to choose papers. *arXiv preprint arxiv:1808.09057* (2018).
[38] K Okike, K Hug, M Kocher, and S Leopold. 2016. Single-blind vs Double-blind Peer Review in the Setting of Author Prestige. *JAMA* (2016).
[39] S. R. Paul. 1981. Bayesian methods for calibration of examiners. *Brit. J. Math. Statist. Psych.* 34, 2 (1981), 213–223.
[40] Edmund S Phelps. 1972. The statistical theory of racism and sexism. *The American economic review* (1972), 659–661.
[41] M Roos, J Rothe, and B Scheuermann. 2011. How to Calibrate the Scores of Biased Reviewers by Quadratic Programming. In *AAAI*.
[42] N Shah, S Balakrishnan, A Guntuboyina, and M Wainwright. 2017. Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Transactions on Information Theory* (2017).
[43] N Shah, S Balakrishnan, and M Wainwright. 2019. Low Permutation-rank Matrices: Structural Properties and Noisy Completion. In *JMLR*.
[44] N Shah, B Tabibian, K Muandet, I Guyon, and U Von Luxburg. 2017. Design and Analysis of the NIPS 2016 Review Process. *JMLR* (2017).
[45] Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. 2016. A permutation-based model for crowd labeling: Optimal estimation and robustness. *arXiv:1606.09632* (2016).
[46] I Stelmakh, N Shah, and A Singh. 2019. On Testing for Biases in Peer Review. In *NeurIPS*.
[47] I Stelmakh, N Shah, and A Singh. 2019. PeerReview4All: Fair and Accurate Reviewer Assignment in Peer Review. In *ALT*.
[48] W Tang, J Tang, and C Tan. 2010. Expertise Matching via Constraint-Based Optimization. In *Conference on Web Intelligence and Intelligent Agent Technology*.
[49] A Tomkins, M Zhang, and W Heavlin. 2017. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences* (2017).
[50] J Wang and N Shah. 2018. Thereś Lots in a Name (Whereas There Shouldńt Be). https://researchonresearch.blog/2018/11/28/theres-lots-in-a-name/.
[51] J Wang and N Shah. 2019. Gender Distributions of Paper Awards. https://researchonresearch.blog/2019/06/18/gender-distributions-of-paper-awards/.
[52] Jingyan Wang and Nihar B Shah. 2019. Your 2 is My 1, Your 3 is My 9: Handling Arbitrary Miscalibrations in Ratings. In *AAMAS*.
[53] T Webb, B OHara, and R Freckleton. 2008. Does double-blind review benefit female authors? *Trends in Ecology and Evolution* (2008).
[54] Y Xu, H Zhao, X Shi, and N Shah. 2018. On Strategyproof Conference Review. In *IJCAI*.
[55] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *ICML*.