shop; discussions from that meeting shaped every aspect of this document. Also, our thanks to Helen Wright, Ann Drobnis, Cynthia Dwork, Sampath Kannan, Michael Kearns, Toni Pitassi, and Suresh Venkatasubramanian. C

References

1. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J. and Wallach, H. A reductions approach to fair classification. In *Proceedings of the 35th Intern. Conf. Machine Learning*. ICML, JMLR Workshop and Conference Proceedings, 2018, 2569–2577.
2. Bastani, H., Bayati, M. and Khosravi, K. Exploiting the natural exploration in contextual bandits. arXiv preprint, 2017, arXiv:1704.09011.
3. Becker, G.S. *The Economics of Discrimination*. University of Chicago Press, 2010.
4. Berk, R., Heidari, H., Jabbari, S., Kearns, M. and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0(0):0049124118782533.
5. Beutel, A., Chen, J., Zhao, Z. and Chi, E.H. Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint, 2017, arXiv:1707.00075.
6. Bird, S., Barocas, S., Crawford, K., Diaz, F. and Wallach, H. Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI. In *Proceedings of Workshop on Fairness, Accountability, and Transparency in Machine Learning*. ACM, 2016.
7. Bolukbasi, T., Chang, K-W., Zou, J.Y., Saligrama, V. and Kalai, A.T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 2016, 4349–4357.
8. Bower, A. et al. Fair pipelines. arXiv preprint, 2017, arXiv:1707.00391.
9. Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. ACM, 2018, 77–91.
10. Calders, T. and Verwer, S. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery 21*, 2 (2010), 277–292.
11. Caliskan, A., Bryson, J.J. and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science 356*, 6334 (2017), 183–186.
12. Celis, L.E., Straszak, D. and Vishnoi, N.K. Ranking with fairness constraints. In *Proceedings of the 45th Intern. Colloquium on Automata, Languages, and Programming*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.
13. Celis, L.E. and Vishnoi, N.K. Fair personalization. arXiv preprint, 2017, arXiv:1707.02260.
14. Chen, I., Johansson, F.D. and Sontag, D. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems*, 2018, 3539–3550.
15. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data 5*, 2 (2017), 153–163.
16. Coat, S. and Loury, G.C. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, 1993, 1220–1240.
17. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*. ACM, 2017, 797¬–806.
18. Doroudi, S., Thomas, P.S. and Brunskill, E. Importance sampling for fair policy selection. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2017.
19. Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conf.* ACM, 2012, 214–226.
20. Dwork, C. and Ilvento, C. Fairness under composition. Manuscript, 2018.
21. Dwork, C., McSherry, F., Nissim, K. and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of Theory of Cryptography Conference*. Springer, 2006, 265–284.
22. Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science 9*, 3–4 (2014), 211–407.
23. Edwards, H. and Storkey, A. Censoring representations with an adversary. arXiv preprint, 2015, arXiv:1511.05897.
24. Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C. and Venkatasubramanian, S. Runaway feedback loops in predictive policing. In *Proceedings of 1st Conf. Fairness, Accountability and Transparency in Computer Science*. ACM, 2018.
25. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C. and Venkatasubramanian, S. Certifying and removing disparate impact. *Proceedings of KDD*, 2015.
26. Foster, D.P. and Vohra, R.A. An economic argument for affirmative action. *Rationality and Society 4*, 2 (1992), 176–188.
27. Friedler, S.A., Scheidegger, C. and Venkatasubramanian, S. On the (im) possibility of fairness. arXiv preprint, 2016, arXiv:1609.07236.
28. Gillen, S., Jung, C., Kearns, M. and Roth, A. Online learning with an unknown fairness metric. *Advances in Neural Information Processing Systems*, 2018.
29. Hardt, M., Price, E. and Srebro, N. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 2016, 3315–3323.
30. Hébert-Johnson, U., Kim, M.P., Reingold, O. and Rothblum, G.N. Calibration for the (computationally identifiable) masses. In *Proceedings of the 35th Intern. Conf. Machine Learning 80*. ICML, JMLR Workshop and Conference Proceedings, 2018, 2569–2577.
31. Hu, L. and Chen, Y. 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*. P.A. Champin, F.L. Gandon, M. Lalmas, and P.G. Ipeirotis, eds. ACM, 2018, 1389–1398.
32. Ja bbari, S., Joseph, M., Kearns, M., Morgenstern, J.H. and Roth, A. Fairness in reinforcement learning. In *Proceedings of the Intern. Conf. Machine Learning*, 2017, 1617–1626.
33. Johndrow, J.E. Lum, K. et al. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics 13*, 1 (2019), 189–220.
34. Joseph, M., Kearns, M., Morgenstern, J.H., Neel, S. and Roth. A. Fair algorithms for infinite and contextual bandits. In *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society*, 2018.
35. Joseph, M., Kearns, M., Morgenstern, J.H. and Roth, A. Fairness in learning: Classic and contextual bandits. *Advances in Neural Information Processing Systems*, 2016, 325–333.
36. Kamishima, T., Akaho, S. and Sakuma, J. Fairness-aware learning through regularization approach. In *Proceedings of the IEEE 11th Intern. Conf. Data Mining Workshops*. IEEE, 2011, 643–650.
37. Kannan, S. et al. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. ACM, 2017, 369–386.
38. Kannan, S., Morgenstern, J., Roth, A., Waggoner, B. and Wu, Z.S. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in Neural Information Processing Systems*, 2018.
39. Kannan, S., Roth, A. and Ziani, J. Downstream effects of affirmative action. In *Proceedings of the Conf. Fairness, Accountability, and Transparency*. ACM, 2019, 240–248.
40. Kearns, M.J., Neel, S., Roth, A. and Wu, Z.S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*. J.G. Dy and A. Krause, eds. JMLR Workshop and Conference Proceedings, ICML, 2018. 2569–2577.
41. Kearns, M., Neel, S., Roth, A. and Wu, Z.S. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conf. Fairness, Accountability, and Transparency*. ACM, 2019, 100–109.
42. Kearns, M., Roth, A. and Wu, Z.S. Meritocratic fairness for cross-population selection. In *Proceedings of International Conference on Machine Learning*, 2017, 1828–1836.
43. Kilbertus, N. et al. Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, 2017, 656–666.
44. Kim, M.P., Ghorbani, A. and Zou, J. Multiaccuracy: Blackbox postprocessing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2019, 247–254.
45. Kim, M.P., Reingold, O. and Rothblum, G.N. Fairness through computationally bounded awareness. *Advances in Neural Information Processing Systems*, 2018.
46. Kleinberg, J.M., Mullainathan, S. and Raghavan,M. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017.
47. Kleinberg, J. and Raghavan, M. Selection problems in the presence of implicit bias. In *Proceedings of the 9th Innovations in Theoretical Computer Science Conference 94*, 2018, 33. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
48. Kusner, M.J., Loftus, J., Russell, C. and Silva, R. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 2017, 4069–4079.
49. Liu, L.T., Dean, S., Rolf, E., Simchowitz, M. and Hardt, M. Delayed impact of fair machine learning. In *Proceedings of the 35th Intern. Conf. Machine Learning*. ICML, 2018.
50. Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D. and Parkes, D.C. Calibrated fairness in bandits. arXiv preprint, 2017, arXiv:1707.01875.
51. Lum, K. and Isaac, W. To predict and serve? *Significance 13*, 5 (2016), 14–19.
52. Madras, D., Creager, E., Pitassi, T. and Zemel, R. Learning adversarially fair and transferable representations. In *Proceedings of Intern. Conf. Machine Learning*, 2018, 3381–3390.
53. Madras, D., Pitassi, T. and Zemel, R.S. Predict responsibly: Increasing fairness by learning to defer. CoRR, 2017, abs/1711.06664.
54. McNamara, D., Ong, C.S. and Williamson, R.C. Provably fair representations. arXiv preprint, 2017, arXiv:1710.04394.
55. Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence 2018* (2018), 1931. NIH Public Access.
56. Pedreshi, D., Ruggieri, S. and Turini, F. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*. ACM, 2008, 560–568.
57. Raghavan, M., Slivkins, A., Wortman Vaughan, J. and Wu, Z.S. The unfair externalities of exploration. *Conference on Learning Theory*, 2018.
58. Rothblum, G.N. and Yona, G. Probably approximately metric-fair learning. In *Proceedings of the 35th Intern. Conf. Machine Learning*. JMLR Workshop and Conference Proceedings, ICML 80 (2018), 2569–2577.
59. Rothwell, J. How the war on drugs damages black social mobility. The Brookings Institution, Sept. 30, 2014.
60. Sweeney, L. Discrimination in online ad delivery. *Queue 11*, 3 (2013), 10.
61. Woodworth, B., Gunasekar, S., Ohannessian, M.I. and Srebro, N. Learning non-discriminatory predictors. In *Proceedings of Conf. Learning Theory*, 2017, 1920–1953.
62. Yang, K. and Stoyanovich, J. Measuring fairness in ranked outputs. In *Proceedings of the 29th Intern. Conf. Scientific and Statistical Database Management*. ACM, 2017, 22.
63. Zafar, M.B., Valera, I. Gomez-Rodriguez, M. and Gummadi, K.P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th Intern. Conf. World Wide Web*. ACM, 2017, 1171–1180.
64. Zemel, R., Wu, Y., Swersky, K., Pitassi, T. and Dwork, C. Learning fair representations. In *Proceedings of ICML*, 2013.
65. Zhang, B.H., Lemoine, B. and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conf. AI, Ethics, and Society*. ACM, 2018, 335–340.

Alexandra Chouldechova (achould@cmu.edu) is Estella Loomis Assistant Professor of Statistics and Public Polict in the Heinz College at Carnegie Mellon University, Pittsburgh, PA, USA.

Aaron Roth (aaroth@cis.upenn.edu) is Class of 1940 Associate Professor in the Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA. Together with Michael Kearns, he is the author of *The Ethical Algorithm*.

Watch the authors discuss this work in the exclusive *Communications* video. https://cacm.acm.org/videos/frontiers-of-fairness

Week 2: Algorithmic Fairness

WE ARE AI
#4

All about that BIAS

TERMS OF USE

All the panels in this comic book are licensed CC BY-NC-ND 4.0. Please refer to the license page for details on how you can use this artwork.

TL;DR: Feel free to use panels/groups of panels in your presentations/articles, as long as you

1. Provide the proper citation
2. Do not make modifications to the individual panels themselves

Cite as:

Julia Stoyanovich and Falaah Arif Khan. "All about that Bias". *We are AI Comics*, Vol 4 (2021)
https://dataresponsibly.github.io/we-are-ai/comics/vol4_en.pdf

Contact:

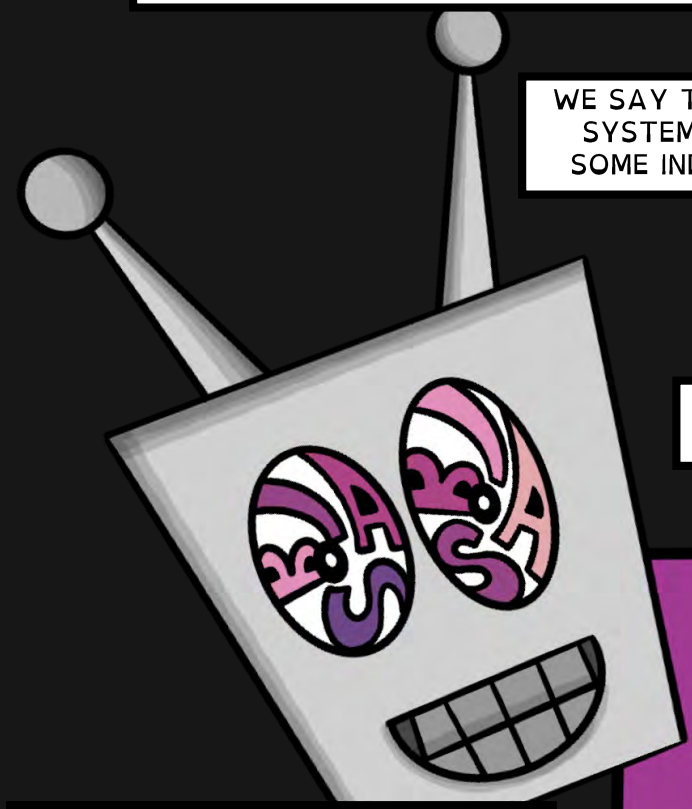
Please direct any queries about using elements from this comic to themachinelearnist@gmail.com and cc stoyanovich@nyu.edu

LET'S TALK ABOUT WHAT WE MEAN BY 'BIAS' IN AI, AND HOW IT ARISES.

WE SAY THAT AN AI IS BIASED IF ITS USE CAN LEAD TO SYSTEMATIC AND UNFAIR DISCRIMINATION AGAINST SOME INDIVIDUALS OR GROUPS IN FAVOR OF OTHERS.

BIAS CAN STEM FROM HARMFUL PATTERNS PICKED UP FROM THE DATA ITSELF,

OR FROM HOW THE ALGORITHM IS DESIGNED,

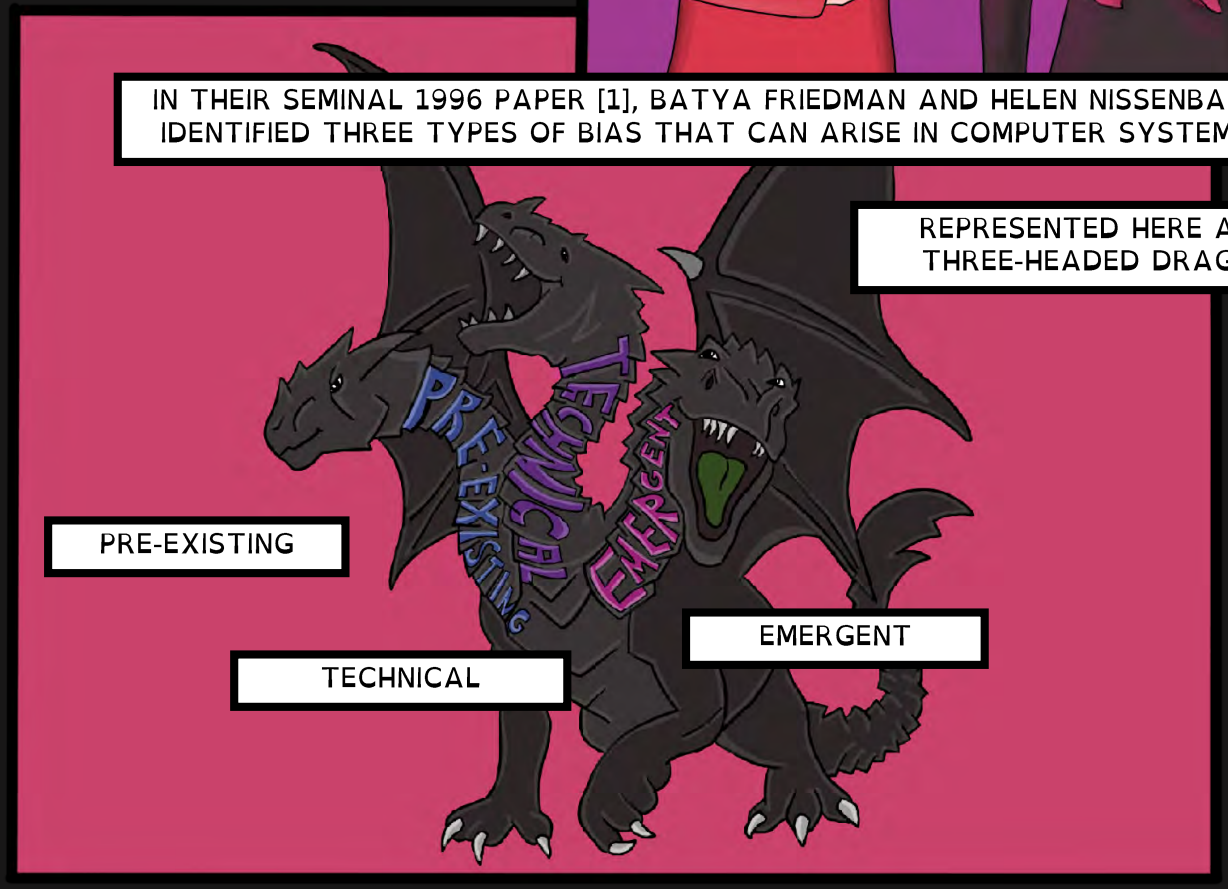


OR FROM THE OBJECTIVES THAT WE SPECIFIED FOR IT,

OR FROM HOW WE USE IT.

IN THEIR SEMINAL 1996 PAPER [1], BATYA FRIEDMAN AND HELEN NISSENBAUM IDENTIFIED THREE TYPES OF BIAS THAT CAN ARISE IN COMPUTER SYSTEMS,

REPRESENTED HERE AS A THREE-HEADED DRAGON:



PRE-EXISTING

TECHNICAL

EMERGENT

[1] Batya Friedman and Helen Nissenbaum. (1996). Bias in computer systems.

RECALL THE BAKING METAPHOR WE USED TO UNDERSTAND DATA-DRIVEN ALGORITHMS IN VOLUME 1.

LET'S NOW USE THE SAME METAPHOR TO UNDERSTAND BIAS!

PRE-EXISTING BIAS EXISTS INDEPENDENT OF THE ALGORITHM AND HAS ITS ORIGINS IN SOCIETY.

THESE WOULD BE THE FLAVOR NOTES THAT WILL SEEP INTO YOUR BREAD IF YOU DON'T PRIORITIZE THE PURITY/FRESHNESS OF YOUR INGREDIENTS,

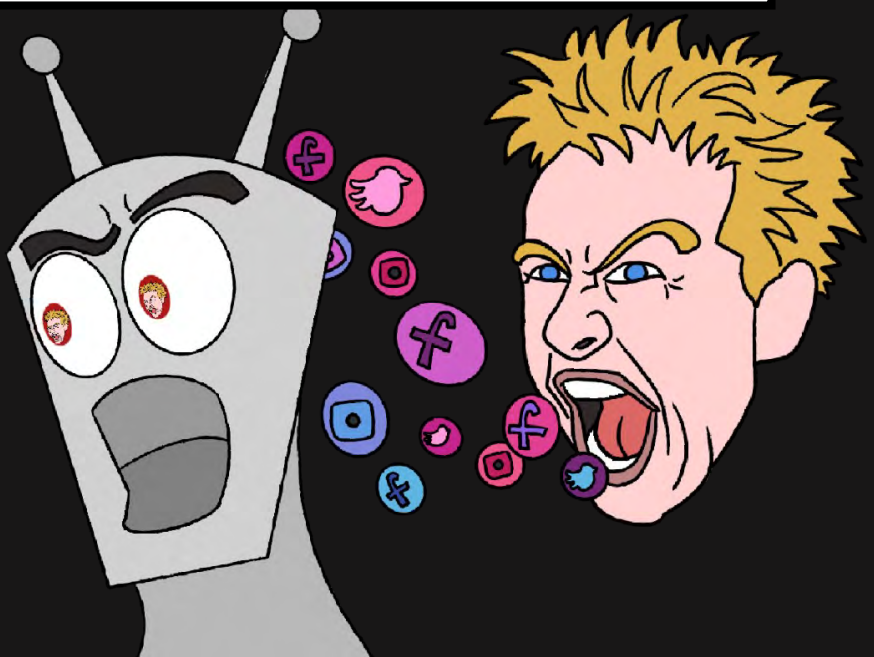
PRE-EXISTING BIAS
(IN THE DATA)

OR IF YOU DECIDE TO USE PREMIXED OFF-THE-SHELF BATTER.

THESE BIASES EXIST IN SOCIETY AND COME 'PRE-BAKED' INTO THE ALGORITHM,

FROM THE UNDERLYING DISCRIMINATORY SYSTEM THAT THE DATA WAS COLLECTED FROM -

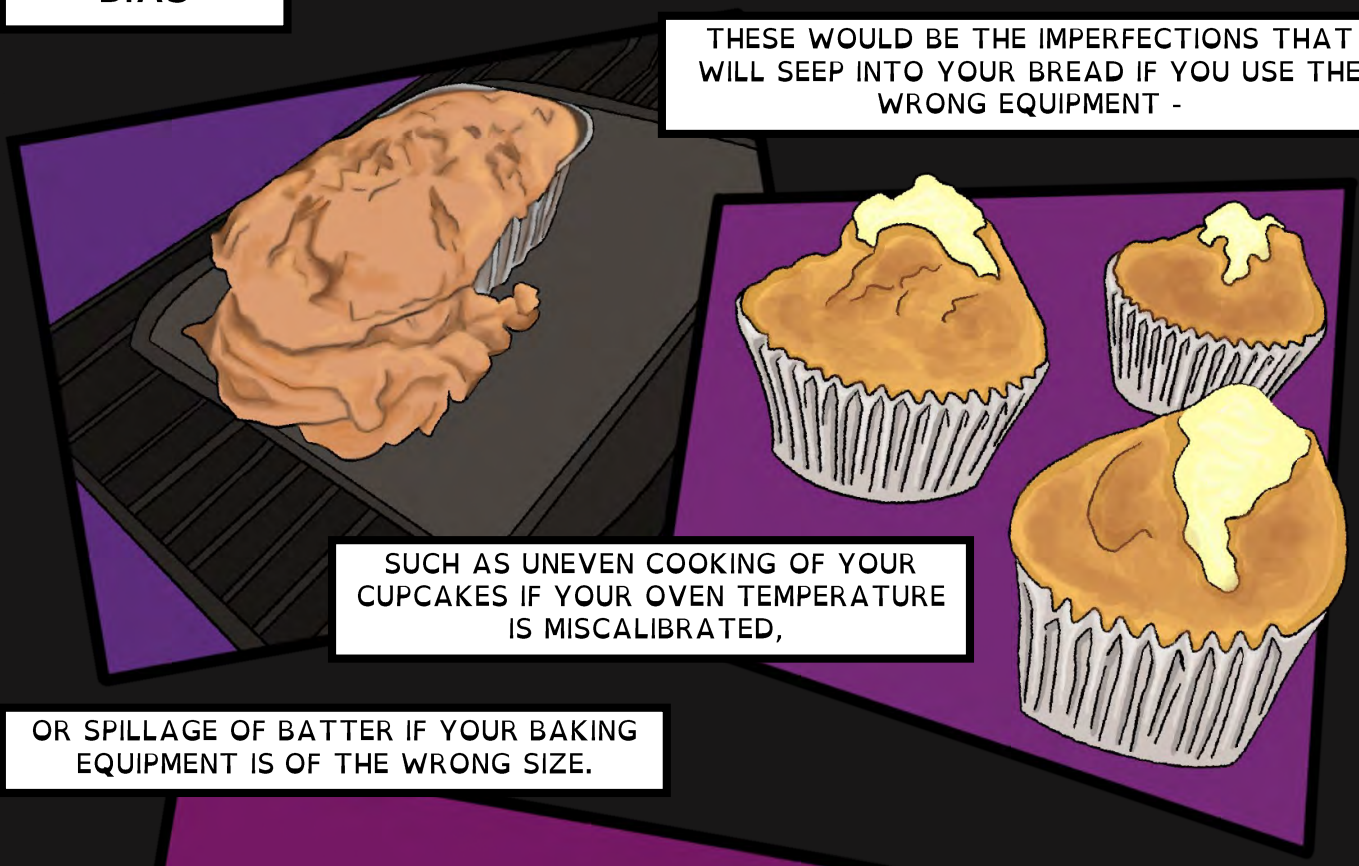
SUCH AS THE GENDER AND RACIAL STEREOTYPES THAT LANGUAGE MODELS PICK UP WHEN TRAINED ON DATA FROM SOCIAL MEDIA.



TECHNICAL BIAS

TECHNICAL BIAS IS INTRODUCED BY THE SYSTEM ITSELF - BECAUSE OF THE WAY IT IS DESIGNED OR OPERATES.

THESE WOULD BE THE IMPERFECTIONS THAT WILL SEEP INTO YOUR BREAD IF YOU USE THE WRONG EQUIPMENT -



SUCH AS UNEVEN COOKING OF YOUR CUPCAKES IF YOUR OVEN TEMPERATURE IS MISCALIBRATED,

OR SPILLAGE OF BATTER IF YOUR BAKING EQUIPMENT IS OF THE WRONG SIZE.

BACK TO COMPUTER SYSTEMS:

A PROMINENT EXAMPLE IS SOCIAL MEDIA PLATFORMS

- DESIGNED TO OPTIMIZE FOR ENGAGEMENT (INSTEAD OF SAFETY OR AUTHENTICITY) -

THAT END UP PROMOTING POLARIZING ARTICLES AND FAKE NEWS.

EMERGENT BIAS
(DUE TO DECISIONS)

EMERGENT BIAS ARISES OVER TIME, BECAUSE THE DECISIONS MADE WITH THE HELP OF THE SYSTEM CHANGE THE WORLD,

WHICH IN TURN IMPACTS THE OPERATION OF THE SYSTEM GOING FORWARD.

THINK ABOUT BEHAVIORAL CHANGES THAT WILL EMERGE AS A RESULT OF YOUR BAKING -

WHAT IF YOU BECOME SUCH A MAESTRO AT BAKING THAT YOU INADVERTENTLY MAKE BREAD A STEADY PART OF YOUR DIET!

OR MAKE IT SO OFTEN, THAT YOU TURN EVERYONE AROUND YOU OFF THE THOUGHT OF EVER EATING ANOTHER SLICE!

OR THINK ABOUT HOW YOUR IDEA OF 'WHAT BREAD SHOULD TASTE LIKE' IS SHAPED BY THE POPULARITY OF PRODUCTS LIKE 'WONDER BREAD'.

IN THE SAME VEIN, THINK ABOUT HOW YOUR EXPOSURE TO NEWS - AND INFORMATION MORE BROADLY -

IS SHAPED BY ALGORITHMS THAT CURATE SOCIAL FEEDS WITH POPULAR AND 'TRENDING' POSTS.

TO MAKE OUR DISCUSSION CONCRETE, LET'S LOOK AT REAL-WORLD EXAMPLES OF ALGORITHMIC BIAS.

LET'S TAKE 'HIRING' AS A REPRESENTATIVE DOMAIN IN WHICH ALGORITHMS ARE INCREASINGLY BEING USED TO MAKE CRITICAL DECISIONS MORE 'EFFICIENTLY'.

ONE OF THE EARLIEST INDICATIONS THAT THERE IS CAUSE FOR CONCERN CAME IN 2015, WITH THE RESULTS OF THE ADFISHER STUDY OUT OF CARNEGIE MELLON UNIVERSITY. [2]

RESEARCHERS RAN AN EXPERIMENT, IN WHICH THEY CREATED TWO SETS OF SYNTHETIC PROFILES OF WEB USERS WHO WERE THE SAME IN EVERY RESPECT

— IN TERMS OF THEIR DEMOGRAPHICS, STATED INTERESTS, AND BROWSING PATTERNS —

WITH A SINGLE EXCEPTION: THEIR STATED GENDER, MALE OR FEMALE.

RESEARCHERS SHOWED THAT GOOGLE DISPLAYED ADS FOR A CAREER COACHING SERVICE FOR HIGH-PAYING EXECUTIVE JOBS FAR MORE FREQUENTLY TO THE MALE GROUP THAN TO THE FEMALE GROUP.

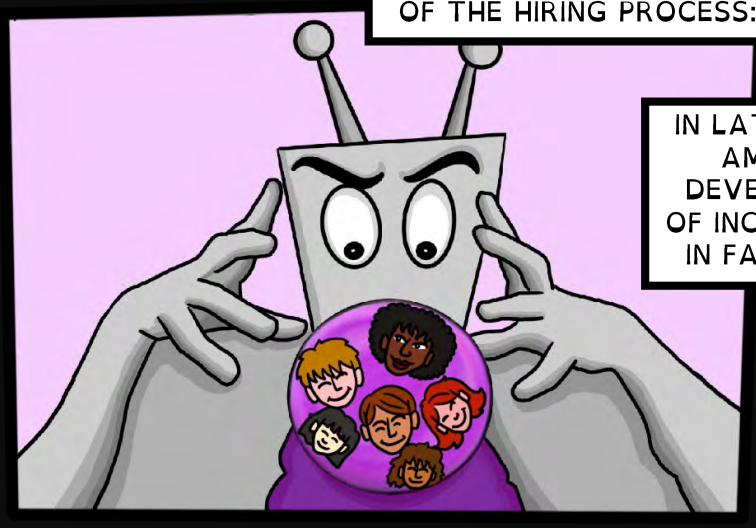
THIS BRINGS BACK MEMORIES OF THE TIME WHEN IT WAS LEGAL TO ADVERTISE JOBS BY GENDER IN NEWSPAPERS. THIS PRACTICE WAS OUTLAWED IN THE US IN 1964, BUT IT PERSISTS IN THE ONLINE AD ENVIRONMENT.

IT WAS LATER SHOWN THAT PART OF THE REASON THIS WAS HAPPENING IS THE MECHANICS OF THE ADVERTISEMENT TARGETING SYSTEM ITSELF, AS AN ARTIFACT OF THE BIDDING PROCESS.

THIS IS TECHNICAL BIAS IN ACTION!

[2] Women less likely to be shown ads for high-paid jobs on Google, study shows. Guardian (2015)

LET US MOVE FORWARD TO THE NEXT STAGE OF THE HIRING PROCESS: **RESUME SCREENING.**



IN LATE 2018 IT WAS REPORTED THAT AMAZON'S AI RECRUITING TOOL, DEVELOPED WITH THE STATED GOAL OF INCREASING WORKFORCE DIVERSITY, IN FACT DID THE OPOSITE THING: [3]

THE SYSTEM TAUGHT ITSELF THAT MALE CANDIDATES WERE PREFERABLE TO FEMALE CANDIDATES.

IT PENALIZED RESUMES THAT INCLUDED THE WORD "WOMEN'S," AS IN "WOMEN'S CHESS CLUB CAPTAIN."

AND IT DOWNGRADED GRADUATES OF TWO ALL-WOMEN'S COLLEGES.

THE RESULTS ALIGNED WITH, AND REINFORCED, A STARK GENDER IMBALANCE IN THE WORKFORCE.

THIS IS EMERGENT BIAS IN ACTION -

A HIRING MANAGER TO WHOM AN AI TOOL REPEATEDLY SUGGEST THE SAME KIND OF JOB APPLICANT AS A GOOD FIT,

WILL OVERTIME COME TO BELIEVE THAT THIS IS WHAT A PROMISING EMPLOYEE LOOKS LIKE.

WE ARE ALSO SEEING PRE-EXISTING BIAS IN THIS EXAMPLE: THE AI TOOL WAS TRAINED ON HISTORICAL DATA ABOUT PAST EMPLOYEES, WHO WERE PREDOMINANTLY MALE

[3] Amazon scraps secret AI recruiting tool that showed bias against women. Reuters (2018)

HERE'S ANOTHER EXAMPLE, LATER YET IN THE HIRING PROCESS, PERHAPS DURING A POST-INTERVIEW BACKGROUND CHECK BY A POTENTIAL EMPLOYER -

LATANYA SWEENEY, A COMPUTER SCIENCE PROFESSOR ON THE FACULTY AT HARVARD,

SHOWED THAT GOOGLING FOR AFRICAN-AMERICAN SOUNDING NAMES IS MORE LIKELY TO TRIGGER ADS SUGGESTIVE OF A CRIMINAL RECORD THAN GOOGLING FOR WHITE-SOUNDING NAMES,

EVEN CONTROLLING FOR WHETHER AN INDIVIDUAL IN FACT HAS A CRIMINAL RECORD! [4]

THIS IS PRE-EXISTING BIAS AT PLAY -

MANIFESTING LONG-STANDING RACIAL PREJUDICES OF SOCIETY.

[4] Racism is Poisoning Online Ad Delivery, Says Harvard Professor. MIT Technology Review (2013)

THE CASES PRESENTED HERE HAVE ONE THING IN COMMON: THEY SHOW THAT AI CAN REINFORCE AND EXACERBATE UNLAWFUL DISCRIMINATION AGAINST MINORITY AND HISTORICALLY DISADVANTAGED GROUPS.

OFTEN THIS IS CALLED OUT AS "BIAS IN AI".

SO, WHY ARE SOPHISTICATED SYSTEMS THAT AIM TO MAKE HIRING MORE EFFICIENT FAILING AT THIS, AND ARGUABLY MAKING THINGS WORSE?

OF COURSE, THE ISSUES OF BIAS IN EMPLOYMENT ARE NOT NEW. THEY EXHIBITED THEMSELVES IN THE ANALOG ERA AS WELL.

FOR EXAMPLE, IN THEIR WELL-KNOWN 2004 STUDY, MARIANNE BERTRAND AND SENDHIL MULLAINATHAN SENT FICTITIOUS RESUMES TO HELP-WANTED ADS IN BOSTON AND CHICAGO NEWSPAPERS. [5]

Are EMILY and GREG more employable than LAKISHA and JAMAL?

TO MANIPULATE PERCEIVED RACE, THEY RANDOMLY ASSIGNED AFRICAN-AMERICAN- OR WHITE-SOUNDING NAMES TO RESUMES.

WHITE NAMES RECEIVE 50 PERCENT MORE CALLBACKS FOR INTERVIEWS.

THIS CASE SHOWS THAT BIAS CAN BE DUE TO HUMAN DECISIONS.

[5] Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. Marianne Bertrand & Sendhil Mullainathan (2003)

LET'S REVISIT PRE-EXISTING BIAS THAT OFTEN EXHIBITS ITSELF IN THE DATA.

DATA IS AN IMAGE OF THE **WORLD**, ITS **MIRROR REFLECTION**.

WHEN WE THINK ABOUT BIAS IN THE DATA, WE INTERROGATE THIS REFLECTION.

ONE INTERPRETATION OF "BIAS IN THE DATA" IS THAT THE REFLECTION IS DISTORTED -

WE MAY SYSTEMATICALLY OVER-REPRESENT OR UNDER-REPRESENT PARTICULAR PARTS OF THE WORLD IN THE DATA,

OR OTHERWISE **DISTORT** THE READINGS.

RECALL THE FAILURE OF AMAZON'S RECRUITING AI TO IMPROVE WORKFORCE DIVERSITY.

THIS TOOL WAS TRAINED USING HISTORICAL DATA: RESUMES OF PEOPLE WHO WERE HIRED IN THE PAST.

THAT TRAINING WAS SUBJECT TO PRE-EXISTING BIAS.

IN THAT DATA, THERE WAS AN UNDER-REPRESENTATION OF WOMEN IN THE WORKFORCE, AND IN TECHNICAL ROLES.

A MORE SUBTLE POINT IS ABOUT DISTORTIONS.

WHEN WE CONSIDER FEATURES, LIKE AN INDIVIDUAL'S SCORE ON A STANDARDIZED TEST, DO WE TAKE THESE AT FACE VALUE?

OR DO WE ACCOUNT FOR DIFFERENCES IN ACCESS TO EDUCATIONAL OPPORTUNITY,

LIKE GOING TO A BETTER SCHOOL, OR HAVING ACCESS TO PAID TUTORING?

ANOTHER INTERPRETATION OF "BIAS IN THE DATA" IS THAT EVEN IF WE WERE ABLE TO REFLECT THE WORLD PERFECTLY IN THE DATA,

IT WOULD STILL BE A REFLECTION OF THE WORLD SUCH AS IT IS,

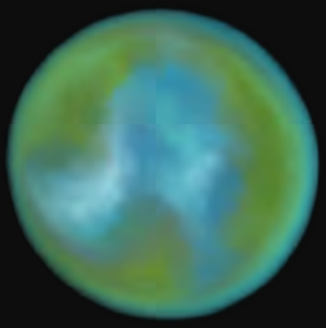
AND NOT NECESSARILY OF HOW IT COULD OR SHOULD BE.

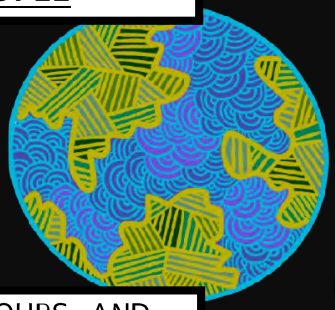
IT IS IMPORTANT TO KEEP IN MIND THAT A REFLECTION CANNOT KNOW WHETHER IT IS DISTORTED.

DATA ALONE CANNOT TELL US WHETHER IT IS A DISTORTED REFLECTION OF A PERFECT WORLD, A PERFECT REFLECTION OF A DISTORTED WORLD,

OR IF THESE DISTORTIONS COMPOUND.

THE SECOND POINT IS THAT IT IS NOT UP TO DATA OR ALGORITHMS, BUT RATHER UP TO **PEOPLE**





— INDIVIDUALS, GROUPS, AND SOCIETY AT LARGE —

TO COME TO CONSENSUS ABOUT WHETHER THE WORLD IS HOW IT SHOULD BE, OR IF IT NEEDS TO BE IMPROVED.

AND, IF SO, HOW WE SHOULD GO ABOUT IMPROVING IT.

THE FINAL POINT HERE IS THAT CHANGING THE REFLECTION MAY NOT CHANGE THE WORLD.

IF THE REFLECTION ITSELF IS USED TO MAKE IMPORTANT DECISIONS -

FOR EXAMPLE, WHOM TO HIRE OR WHAT SALARY TO OFFER TO AN INDIVIDUAL BEING HIRED,

THEN COMPENSATING FOR THE DISTORTIONS IS WORTHWHILE.

BUT THE MIRROR METAPHOR ONLY TAKES US SO FAR.

WE HAVE TO WORK MUCH HARDER — USUALLY GOING FAR BEYOND TECHNOLOGICAL SOLUTIONS — TO MAKE LASTING CHANGE IN THE WORLD,

NOT MERELY BRUSH UP THE REFLECTION.

CIRCLING BACK NOW TO THE THREE-HEADED BIAS DRAGON.

WHEN SPEAKING ABOUT TACKLING BIAS IN AI, WE TEND TO FRAME THE PROBLEM AS FINDING A WAY TO SLAY THE BIAS-DRAGON.

PRE-EXISTING

TECHNICAL

EMERGENT

BUT THROUGH OUR DISCUSSION OF THE LINK BETWEEN HUMAN BIAS AND MACHINE BIAS,

WE FIND OURSELVES QUESTIONING THE VERY NATURE OF THIS TALE -

AT THE END OF THE DAY, MAYBE THE QUESTION ISN'T -

HOW TO SLAY THE DRAGON AND RESCUE THE PRINCESS?

THE QUESTION WE REALLY SHOULD BE ASKING OURSELVES IS -

WHAT DO WE DO ABOUT A SOCIETY THAT LOCKS UP PRINCESSES IN CASTLES, IN THE FIRST PLACE?

FIN.

Bias in Computer Systems

BATYA FRIEDMAN
Colby College and The Mina Institute
and
HELEN NISSENBAUM
Princeton University

From an analysis of actual cases, three categories of bias in computer systems have been developed: preexisting, technical, and emergent. Preexisting bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias arises in a context of use. Although others have pointed to bias in particular computer systems and have noted the general problem, we know of no comparable work that examines this phenomenon comprehensively and which offers a framework for understanding and remedying it. We conclude by suggesting that freedom from bias should be counted among the select set of criteria—including reliability, accuracy, and efficiency—according to which the quality of systems in use in society should be judged.

Categories and Subject Descriptors: D.2.0 [**Software**]: Software Engineering; H.1.2 [**Information Systems**]: User/Machine Systems; K.4.0 [**Computers and Society**]: General

General Terms: Design, Human Factors

Additional Key Words and Phrases: Bias, computer ethics, computers and society, design methods, ethics, human values, standards, social computing, social impact, system design, universal design, values

INTRODUCTION

To introduce what bias in computer systems might look like, consider the case of computerized airline reservation systems, which are used widely by travel agents to identify and reserve airline flights for their customers. These reservation systems seem straightforward. When a travel agent types in a customer's travel requirements, the reservation system searches

This research was funded in part by the Clare Boothe Luce Foundation.
Earlier aspects of this work were presented at the 4S/EASST Conference, Goteborg, Sweden, August 1992, and at InterCHI '93, Amsterdam, April 1993. An earlier version of this article appeared as Tech. Rep. CSLI-94-188, CSLI, Stanford University.
Authors' addresses: B. Friedman, Department of Mathematics and Computer Science, Colby College, Waterville, ME 04901; email: b_friedm@colby.edu; H. Nissenbaum, University Center for Human Values, Marx Hall, Princeton University, Princeton, NJ 08544; email: helen@phoenix.princeton.edu.

a database of flights and retrieves all reasonable flight options that meet or come close to the customer's requirements. These options then are ranked according to various criteria, giving priority to nonstop flights, more direct routes, and minimal total travel time. The ranked flight options are displayed for the travel agent. In the 1980s, however, most of the airlines brought before the Antitrust Division of the United States Justice Department allegations of anticompetitive practices by American and United Airlines whose reservation systems—Sabre and Apollo, respectively—dominated the field. It was claimed, among other things, that the two reservations systems are biased [Schrifin 1985].

One source of this alleged bias lies in Sabre's and Apollo's algorithms for controlling search and display functions. In the algorithms, preference is given to "on-line" flights, that is, flights with all segments on a single carrier. Imagine, then, a traveler who originates in Phoenix and flies the first segment of a round-trip overseas journey to London on American Airlines, changing planes in New York. All other things being equal, the British Airlines' flight from New York to London would be ranked lower than the American Airlines' flight from New York to London even though in both cases a traveler is similarly inconvenienced by changing planes and checking through customs. Thus, the computer systems systematically downgrade and, hence, are biased against international carriers who fly few, if any, internal U.S. flights, and against internal carriers who do not fly international flights [Fotos 1988; Ott 1988].

Critics also have been concerned with two other problems. One is that the interface design compounds the bias in the reservation systems. Lists of ranked flight options are displayed screen by screen. Each screen displays only two to five options. The advantage to a carrier of having its flights shown on the first screen is enormous since 90% of the tickets booked by travel agents are booked by the first screen display [Taib 1990]. Even if the biased algorithm and interface give only a small percent advantage overall to one airline, it can make the difference to its competitors between survival and bankruptcy. A second problem arises from the travelers' perspective. When travelers contract with an independent third party—a travel agent—to determine travel plans, travelers have good reason to assume they are being informed accurately of their travel options; in many situations, that does not happen.

As Sabre and Apollo illustrate, biases in computer systems can be difficult to identify let alone remedy because of the way the technology engages and extenuates them. Computer systems, for instance, are comparatively inexpensive to disseminate, and thus, once developed, a biased system has the potential for widespread impact. If the system becomes a standard in the field, the bias becomes pervasive. If the system is complex, and most are, biases can remain hidden in the code, difficult to pinpoint or explicate, and not necessarily disclosed to users or their clients. Furthermore, unlike in our dealings with biased individuals with whom a potential victim can negotiate, biased systems offer no equivalent means for appeal.

Although others have pointed to bias in particular computer systems and have noted the general problem [Johnson and Mulvey 1993; Moor 1985], we know of no comparable work that focuses exclusively on this phenomenon and examines it comprehensively.

In this article, we provide a framework for understanding bias in computer systems. From an analysis of actual computer systems, we have developed three categories: preexisting bias, technical bias, and emergent bias. Preexisting bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias arises in a context of use. We begin by defining bias and explicating each category and then move to case studies. We conclude with remarks about how bias in computer systems can be remedied.

1. WHAT IS A BIASED COMPUTER SYSTEM?

In its most general sense, the term bias means simply "slant." Given this undifferentiated usage, at times the term is applied with relatively neutral content. A grocery shopper, for example, can be "biased" by not buying damaged fruit. At other times, the term bias is applied with significant moral meaning. An employer, for example, can be "biased" by refusing to hire minorities. In this article we focus on instances of the latter, for if one wants to develop criteria for judging the quality of systems in use—which we do—then criteria must be delineated in ways that speak robustly yet precisely to relevant social matters. Focusing on bias of moral import does just that.

Accordingly, we use the term bias to refer to computer systems that *systematically* and *unfairly discriminate* against certain individuals or groups of individuals in favor of others. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate. Consider, for example, an automated credit advisor that assists in the decision of whether or not to extend credit to a particular applicant. If the advisor denies credit to individuals with consistently poor payment records we do not judge the system to be biased because it is reasonable and appropriate for a credit company to want to avoid extending credit privileges to people who consistently do not pay their bills. In contrast, a credit advisor that systematically assigns poor credit ratings to individuals with ethnic surnames discriminates on grounds that are not relevant to credit assessments and, hence, discriminates unfairly.

Two points follow. First, unfair discrimination alone does not give rise to bias unless it occurs systematically. Consider again the automated credit advisor. Imagine a random glitch in the system which changes in an isolated case information in a copy of the credit record for an applicant who happens to have an ethnic surname. The change in information causes a downgrading of this applicant's rating. While this applicant experiences unfair discrimination resulting from this random glitch, the applicant could have been anybody. In a repeat incident, the same applicant or others with

similar ethnicity would not be in a special position to be singled out. Thus, while the system is prone to random error, it is not biased.

Second, systematic discrimination does not establish bias unless it is joined with an unfair outcome. A case in point is the Persian Gulf War, where United States Patriot missiles were used to detect and intercept Iraqi Scud missiles. At least one software error identified during the war contributed to systematically poor performance by the Patriots [Gao 1992]. Calculations used to predict the location of a Scud depended in complex ways on the Patriots' internal clock. The longer the Patriot's continuous running time, the greater the imprecision in the calculation. The deaths of at least 28 Americans in Dhahran can be traced to this software error, which systematically degraded the accuracy of Patriot missiles. While we are not minimizing the serious consequence of this systematic computer error, it falls outside of our analysis because it does not involve unfairness.

2. FRAMEWORK FOR ANALYZING BIAS IN COMPUTER SYSTEMS

We derived our framework by examining actual computer systems for bias. Instances of bias were identified and characterized according to their source, and then the characterizations were generalized to more abstract categories. These categories were further refined by their application to other instances of bias in the same or additional computer systems. In most cases, our knowledge of particular systems came from the published literature. In total, we examined 17 computer systems from diverse fields including banking, commerce, computer science, education, medicine, and law.

The framework that emerged from this methodology is comprised of three overarching categories—preexisting bias, technical bias, and emergent bias. Table I contains a detailed description of each category. In more general terms, they can be described as follows.

2.1 Preexisting Bias

Preexisting bias has its roots in social institutions, practices, and attitudes. When computer systems embody biases that exist independently, and usually prior to the creation of the system, then we say that the system embodies preexisting bias. Preexisting biases may originate in society at large, in subcultures, and in formal or informal, private or public organizations and institutions. They can also reflect the personal biases of individuals who have significant input into the design of the system, such as the client or system designer. This type of bias can enter a system either through the explicit and conscious efforts of individuals or institutions, or implicitly and unconsciously, even in spite of the best of intentions. For example, imagine an expert system that advises on loan applications. In determining an applicant's credit risk, the automated loan advisor negatively weights applicants who live in "undesirable" locations, such as low-income or high-crime neighborhoods, as indicated by their home addresses (a practice referred to as "red-lining"). To the extent the program

Table I. Categories of Bias in Computer System Design

These categories describe ways in which bias can arise in the design of computer systems. The illustrative examples portray plausible cases of bias.

1. Preexisting Bias
Preexisting bias has its roots in social institutions, practices, and attitudes.
When computer systems embody biases that exist independently, and usually prior to the creation of the system, then the system exemplifies preexisting bias. Preexisting bias can enter a system either through the explicit and conscious efforts of individuals or institutions, or implicitly and unconsciously, even in spite of the best of intentions.

1.1. Individual
Bias that originates from individuals who have significant input into the design of the system, such as the client commissioning the design or the system designer (e.g., a client embeds personal racial biases into the specifications for loan approval software).

1.2 Societal
Bias that originates from society at large, such as from organizations (e.g., industry), institutions (e.g., legal systems), or culture at large (e.g., gender biases present in the larger society that lead to the development of educational software that overall appeals more to boys than girls).

2. Technical Bias
Technical bias arises from technical constraints or technical considerations.

2.1 Computer Tools
Bias that originates from a limitation of the computer technology including hardware, software, and peripherals (e.g., in a database for matching organ donors with potential transplant recipients certain individuals retrieved and displayed on initial screens are favored systematically for a match over individuals displayed on later screens).

2.2 Decontextualized Algorithms
Bias that originates from the use of an algorithm that fails to treat all groups fairly under all significant conditions (e.g., a scheduling algorithm that schedules airplanes for take-off relies on the alphabetic listing of the airlines to rank order flights ready within a given period of time).

2.3 Random Number Generation
Bias that originates from imperfections in pseudorandom number generation or in the misuse of pseudorandom numbers (e.g., an imperfection in a random-number generator used to select recipients for a scarce drug leads systematically to favoring individuals toward the end of the database).

2.4 Formalization of Human Constructs
Bias that originates from attempts to make human constructs such as discourse, judgments, or intuitions amenable to computers: when we quantify the qualitative, discretize the continuous, or formalize the nonformal (e.g., a legal expert system advises defendants on whether or not to plea bargain by assuming that law can be spelled out in an unambiguous manner that is not subject to human and humane interpretations in context).

Table I. *Continued*

These categories describe ways in which bias can arise in the design of computer systems. The illustrative examples portray plausible cases of bias.
3. Emergent Bias Emergent bias arises in a context of use with real users. This bias typically emerges some time after a design is completed, as a result of changing societal knowledge, population, or cultural values. User interfaces are likely to be particularly prone to emergent bias because interfaces by design seek to reflect the capacities, character, and habits of prospective users. Thus, a shift in context of use may well create difficulties for a new set of users.
3.1 New Societal Knowledge Bias that originates from the emergence of new knowledge in society that cannot be or is not incorporated into the system design (e.g., a medical expert system for AIDS patients has no mechanism for incorporating cutting-edge medical discoveries that affect how individuals with certain symptoms should be treated).
3.2 Mismatch between Users and System Design Bias that originates when the population using the system differs on some significant dimension from the population assumed as users in the design.
3.2.1 Different Expertise Bias that originates when the system is used by a population with a different knowledge base from that assumed in the design (e.g., an ATM with an interface that makes extensive use of written instructions—"place the card, magnetic tape side down, in the slot to your left"—is installed in a neighborhood with primarily a nonliterate population).
3.2.2 Different Values Bias that originates when the system is used by a population with different values than those assumed in the design (e.g., educational software to teach mathematics concepts is embedded in a game situation that rewards individualistic and competitive strategies, but is used by students with a cultural background that largely eschews competition and instead promotes cooperative endeavors).

embeds the biases of clients or designers who seek to avoid certain applicants on the basis of group stereotypes, the automated loan advisor's bias is preexisting.

2.2 Technical Bias

In contrast to preexisting bias, technical bias arises from the resolution of issues in the technical design. Sources of technical bias can be found in several aspects of the design process, including limitations of computer tools such as hardware, software, and peripherals; the process of ascribing social meaning to algorithms developed out of context; imperfections in pseudorandom number generation; and the attempt to make human constructs amenable to computers, when we quantify the qualitative, discretize the continuous, or formalize the nonformal. As an illustration, consider again the case of Sabre and Apollo described above. A technical constraint imposed by the size of the monitor screen forces a piecemeal presentation of flight options and, thus, makes the algorithm chosen to

rank flight options critically important. Whatever ranking algorithm is used, if it systematically places certain airlines' flights on initial screens and other airlines' flights on later screens, the system will exhibit technical bias.

2.3 Emergent Bias

While it is almost always possible to identify preexisting bias and technical bias in a system design at the time of creation or implementation, emergent bias arises only in a context of use. This bias typically emerges some time after a design is completed, as a result of changing societal knowledge, population, or cultural values. Using the example of an automated airline reservation system, envision a hypothetical system designed for a group of airlines all of whom serve national routes. Consider what might occur if that system was extended to include international airlines. A flight-ranking algorithm that favors on-line flights when applied in the original context with national airlines leads to no systematic unfairness. However, in the new context with international airlines, the automated system would place these airlines at a disadvantage and, thus, comprise a case of emergent bias. User interfaces are likely to be particularly prone to emergent bias because interfaces by design seek to reflect the capacities, character, and habits of prospective users. Thus, a shift in context of use may well create difficulties for a new set of users.

3. APPLICATIONS OF THE FRAMEWORK

We now analyze actual computer systems in terms of the framework introduced above. It should be understood that the systems we analyze are by and large good ones, and our intention is not to undermine their integrity. Rather, our intention is to develop the framework, show how it can identify and clarify our understanding of bias in computer systems, and establish its robustness through real-world cases.

3.1 The National Resident Match Program (NRMP)

The NRMP implements a centralized method for assigning medical school graduates their first employment following graduation. The centralized method of assigning medical students to hospital programs arose in the 1950s in response to the chaotic job placement process and on-going failure of hospitals and students to arrive at optimal placements. During this early period the matching was carried out by a mechanical card-sorting process, but in 1974 electronic data processing was introduced to handle the entire matching process. (For a history of the NRMP, see Graettinger and Peranson [1981a].) After reviewing applications and interviewing students, hospital programs submit to the centralized program their ranked list of students. Students do the same for hospital programs. Hospitals and students are not permitted to make other arrangements with one another or to attempt to directly influence each others' rankings prior to the match.

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg * Sendhil Mullainathan † Manish Raghavan ‡

Abstract

Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

1 Introduction

There are many settings in which a sequence of people comes before a decision-maker, who must make a judgment about each based on some observable set of features. Across a range of applications, these judgments are being carried out by an increasingly wide spectrum of approaches ranging from human expertise to algorithmic and statistical frameworks, as well as various combinations of these approaches.

Along with these developments, a growing line of work has asked how we should reason about issues of bias and discrimination in settings where these algorithmic and statistical techniques, trained on large datasets of past instances, play a significant role in the outcome. Let us consider three examples where such issues arise, both to illustrate the range of relevant contexts, and to surface some of the challenges.

A set of example domains. First, at various points in the criminal justice system, including decisions about bail, sentencing, or parole, an officer of the court may use quantitative *risk tools* to assess a defendant's probability of recidivism — future arrest — based on their past history and other attributes. Several recent analyses have asked whether such tools are mitigating or exacerbating the sources of bias in the criminal justice system; in one widely-publicized report, Angwin et al. analyzed a commonly used statistical method for assigning risk scores in the criminal justice system — the COMPAS risk tool — and argued that it was biased against African-American defendants [2, 23]. One of their main contentions was that the tool's errors were asymmetric: African-American defendants were more likely to be incorrectly labeled as higher-risk than they actually were, while white defendants were more likely to be incorrectly labeled as lower-risk than they actually were. Subsequent analyses raised methodological objections to this report, and also observed that despite the COMPAS risk tool's errors, its estimates of the probability of recidivism are equally well calibrated to the true outcomes for both African-American and white defendants [1, 10, 13, 17].

*Cornell University
†Harvard University
‡Cornell University

Second, in a very different domain, researchers have begun to analyze the ways in which different genders and racial groups experience advertising and commercial content on the Internet differently [9, 26]. We could ask, for example: if a male user and female user are equally interested in a particular product, does it follow that they're equally likely to be shown an ad for it? Sometimes this concern may have broader implications, for example if women in aggregate are shown ads for lower-paying jobs. Other times, it may represent a clash with a user's leisure interests: if a female user interacting with an advertising platform is interested in an activity that tends to have a male-dominated viewership, like professional football, is the platform as likely to show her an ad for football as it is to show such an ad to an interested male user?

A third domain, again quite different from the previous two, is medical testing and diagnosis. Doctors making decisions about a patient's treatment may rely on tests providing probability estimates for different diseases and conditions. Here too we can ask whether such decision-making is being applied uniformly across different groups of patients [16, 27], and in particular how medical tests may play a differential role for conditions that vary widely in frequency between these groups.

Providing guarantees for decision procedures. One can raise analogous questions in many other domains of fundamental importance, including decisions about hiring, lending, or school admissions [24], but we will focus on the three examples above for the purposes of this discussion. In these three example domains, a few structural commonalities stand out. First, the algorithmic estimates are often being used as "input" to a larger framework that makes the overall decision — a risk score provided to a human expert in the legal and medical instances, and the output of a machine-learning algorithm provided to a larger advertising platform in the case of Internet ads. Second, the underlying task is generally about classifying whether people possess some relevant property: recidivism, a medical condition, or interest in a product. We will refer to people as being *positive instances* if they truly possess the property, and *negative instances* if they do not. Finally, the algorithmic estimates being provided for these questions are generally not pure yes-no decisions, but instead probability estimates about whether people constitute positive or negative instances.

Let us suppose that we are concerned about how our decision procedure might operate differentially between two groups of interest (such as African-American and white defendants, or male and female users of an advertising system). What sorts of guarantees should we ask for as protection against potential bias?

A first basic goal in this literature is that the probability estimates provided by the algorithm should be *well-calibrated*: if the algorithm identifies a set of people as having a probability z of constituting positive instances, then approximately a z fraction of this set should indeed be positive instances [8, 14]. Moreover, this condition should hold when applied separately in each group as well [13]. For example, if we are thinking in terms of potential differences between outcomes for men and women, this means requiring that a z fraction of men and a z fraction of women assigned a probability z should possess the property in question.

A second goal focuses on the people who constitute positive instances (even if the algorithm can only imperfectly recognize them): the average score received by people constituting positive instances should be the same in each group. We could think of this as *balance for the positive class*, since a violation of it would mean that people constituting positive instances in one group receive consistently lower probability estimates than people constituting positive instances in another group. In our initial criminal justice example, for instance, one of the concerns raised was that white defendants who went on to commit future crimes were assigned risk scores corresponding to lower probability estimates in aggregate; this is a violation of the condition here. There is a completely analogous property with respect to negative instances, which we could call *balance for the negative class*. These balance conditions can be viewed as generalizations of the notions that both groups should have equal false negative and false positive rates.

It is important to note that balance for the positive and negative classes, as defined here, is distinct in

crucial ways from the requirement that the average probability estimate globally over *all* members of the two groups be equal. This latter global requirement is a version of *statistical parity* [12, 4, 21, 22]. In some cases statistical parity is a central goal (and in some it is legally mandated), but the examples considered so far suggest that classification and risk assessment are much broader activities where statistical parity is often neither feasible nor desirable. Balance for the positive and negative classes, however, is a goal that can be discussed independently of statistical parity, since these two balance conditions simply ask that once we condition on the "correct" answer for a person, the chance of making a mistake on them should not depend on which group they belong to.

The present work: Trade-offs among the guarantees. Despite their different formulations, the calibration condition and the balance conditions for the positive and negative classes intuitively all seem to be asking for variants of the same general goal — that our probability estimates should have the same effectiveness regardless of group membership. One might therefore hope that it would be feasible to achieve all of them simultaneously.

Our main result, however, is that these conditions are in general incompatible with each other; they can only be simultaneously satisfied in certain highly constrained cases. Moreover, this incompatibility applies to *approximate* versions of the conditions as well.

In the remainder of this section we formulate this main result precisely, as a theorem building on a model that makes the discussion thus far more concrete.

1.1 Formulating the Goal

Let's start with some basic definitions. As above, we have a collection of people each of whom constitutes either a positive instance or a negative instance of the classification problem. We'll say that the *positive class* consists of the people who constitute positive instances, and the negative class consists of the people who constitute negative instances. For example, for criminal defendants, the positive class could consist of those defendants who will be arrested again within some fixed time window, and the negative class could consist of those who will not. The positive and negative classes thus represent the "correct" answer to the classification problem; our decision procedure does not know them, but is trying to estimate them.

Feature vectors. Each person has an associated *feature vector* σ, representing the data that we know about them. Let p_σ denote the fraction of people with feature vector σ who belong to the positive class. Conceptually, we will picture that while there is variation within the set of people who have feature vector σ, this variation is invisible to whatever decision procedure we apply; all people with feature vector σ are indistinguishable to the procedure. Our model will assume that the value p_σ for each σ is known to the procedure.[1]

Groups. Each person also belongs to one of two *groups*, labeled 1 or 2, and we would like our decisions to be unbiased with respect to the members of these two groups.[2] In our examples, the two groups could correspond to different races or genders, or other cases where we want to look for the possibility of bias between them. The two groups have different distributions over feature vectors: a person of group t has a probability $a_{t\sigma}$ of exhibiting the feature vector σ. However, people of each group have the same probability

[1] Clearly the case in which the value of p_σ is unknown is an important version of the problem as well; however, since our main results establish strong limitations on what is achievable, these limitations are only stronger because they apply even to the case of known p_σ.

[2] We focus on the case of two groups for simplicity of exposition, but it is straightforward to extend all of our definitions to the case of more than two groups.

p_σ of belonging to the positive class provided their feature vector is σ. In this respect, σ contains all the relevant information available to us about the person's future behavior; once we know σ, we do not get any additional information from knowing their group as well.[3]

Risk Assignments. We say that an *instance* of our problem is specified by the parameters above: a feature vector and a group for each person, with a value p_σ for each feature vector, and distributions $\{a_{t\sigma}\}$ giving the frequency of the feature vectors in each group.

Informally, risk assessments are ways of dividing people up into sets based on their feature vectors σ (potentially using randomization), and then assigning each set a probability estimate that the people in this set belong to the positive class. Thus, we define a *risk assignment* to consist of a set of "bins" (the sets), where each bin is labeled with a *score* v_b that we intend to use as the probability for everyone assigned to bin b. We then create a rule for assigning people to bins based on their feature vector σ; we allow the rule to divide people with a fixed feature vector σ across multiple bins (reflecting the possible use of randomization). Thus, the rule is specified by values $X_{\sigma b}$: a fraction $X_{\sigma b}$ of all people with feature vector σ are assigned to bin b. Note that the rule does not have access to the group t of the person being considered, only their feature vector σ. (As we will see, this does not mean that the rule is incapable of exhibiting bias between the two groups.) In summary, a risk assignment is specified by a set of bins, a score for each bin, and values $X_{\sigma b}$ that define a mapping from people with feature vectors to bins.

Fairness Properties for Risk Assignments. Within the model, we now express the three conditions discussed at the outset, each reflecting a potentially different notion of what it means for the risk assignment to be "fair."

(A) *Calibration within groups* requires that for each group t, and each bin b with associated score v_b, the expected number of people from group t in b who belong to the positive class should be a v_b fraction of the expected number of people from group t assigned to b.

(B) *Balance for the negative class* requires that the average score assigned to people of group 1 who belong to the negative class should be the same as the average score assigned to people of group 2 who belong to the negative class. In other words, the assignment of scores shouldn't be systematically more inaccurate for negative instances in one group than the other.

(C) *Balance for the positive class* symmetrically requires that the average score assigned to people of group 1 who belong to the positive class should be the same as the average score assigned to people of group 2 who belong to the positive class.

Why Do These Conditions Correspond to Notions of Fairness?. All of these are natural conditions to impose on a risk assignment; and as indicated by the discussion above, all of them have been proposed as versions of fairness. The first one essentially asks that the scores mean what they claim to mean, even when considered separately in each group. In particular, suppose a set of scores lack the first property for some bin b, and these scores are given to a decision-maker; then if people of two different groups both belong to bin b, the decision-maker has a clear incentive to treat them differently, since the lack of calibration within groups on bin b means that these people have different aggregate probabilities of belonging to the positive class. Another way of stating the property of calibration within groups is to say that, conditioned on the bin to which an individual is assigned, the likelihood that the individual is a member of the positive class is independent of the group to which the individual belongs. This means we are justified in treating people

[3] As we will discuss in more detail below, the assumption that the group provides no additional information beyond σ does not restrict the generality of the model, since we can always consider instances in which people of different groups never have the same feature vector σ, and hence σ implicitly conveys perfect information about a person's group.

with the same score comparably with respect to the outcome, rather than treating people with the same score differently based on the group they belong to.

The second and third ask that if two individuals in different groups exhibit comparable future behavior (negative or positive), they should be treated comparably by the procedure. In other words, a violation of, say, the second condition would correspond to the members of the negative class in one group receiving consistently higher scores than the members of the negative class in the other group, despite the fact that the members of the negative class in the higher-scoring group have done nothing to warrant these higher scores.

We can also interpret some of the prior work around our earlier examples through the lens of these conditions. For example, in the analysis of the COMPAS risk tool for criminal defendants, the critique by Angwin et al. focused on the risk tool's violation of conditions (B) and (C); the counter-arguments established that it satisfies condition (A). While it is clearly crucial for a risk tool to satisfy (A), it may still be important to know that it violates (B) and (C). Similarly, to think in terms of the example of Internet advertising, with male and female users as the two groups, condition (A) as before requires that our estimates of ad-click probability mean the same thing in aggregate for men and women. Conditions (B) and (C) are distinct; condition (C), for example, says that a female user who genuinely wants to see a given ad should be assigned the same probability as a male user who wants to see the ad.

1.2 Determining What is Achievable: A Characterization Theorem

When can conditions (A), (B), and (C) be simultaneously achieved? We begin with two simple cases where it's possible.

- *Perfect prediction*. Suppose that for each feature vector σ, we have either $p_\sigma = 0$ or $p_\sigma = 1$. This means that we can achieve perfect prediction, since we know each person's class label (positive or negative) for certain. In this case, we can assign all feature vectors σ with $p_\sigma = 0$ to a bin b with score $v_b = 0$, and all σ with $p_\sigma = 1$ to a bin b' with score $v_{b'} = 1$. It is easy to check that all three of the conditions (A), (B), and (C) are satisfied by this risk assignment.
- *Equal base rates*. Suppose, alternately, that the two groups have the same fraction of members in the positive class; that is, the average value of p_σ is the same for the members of group 1 and group 2. (We can refer to this as the *base rate* of the group with respect to the classification problem.) In this case, we can create a single bin b with score equal to this average value of p_σ, and we can assign everyone to bin b. While this is not a particularly informative risk assignment, it is again easy to check that it satisfies fairness conditions (A), (B), and (C).

Our first main result establishes that these are in fact the only two cases in which a risk assignment can achieve all three fairness guarantees simultaneously.

Theorem 1.1 *Consider an instance of the problem in which there is a risk assignment satisfying fairness conditions (A), (B), and (C). Then the instance must either allow for perfect prediction (with p_σ equal to 0 or 1 for all σ) or have equal base rates.*

Thus, in every instance that is more complex than the two cases noted above, there will be some natural fairness condition that is violated by any risk assignment. Moreover, note that this result applies regardless of how the risk assignment is computed; since our framework considers risk assignments to be arbitrary functions from feature vectors to bins labeled with probability estimates, it applies independently of the method — algorithmic or otherwise — that is used to construct the risk assignment.

The conclusions of the first theorem can be relaxed in a continuous fashion when the fairness conditions are only approximate. In particular, for any $\varepsilon > 0$ we can define ε-approximate versions of each of conditions (A), (B), and (C) (specified precisely in the next section), each of which requires that the corresponding equalities between groups hold only to within an error of ε. For any $\delta > 0$, we can also define a δ-approximate version of the equal base rates condition (requiring that the base rates of the two groups be within an additive δ of each other) and a δ-approximate version of the perfect prediction condition (requiring that in each group, the average of the expected scores assigned to members of the positive class is at least $1 - \delta$; by the calibration condition, this can be shown to imply a complementary bound on the average of the expected scores assigned to members of the negative class).

In these terms, our approximate version of Theorem 1.1 is the following.

Theorem 1.2 *There is a continuous function f, with $f(x)$ going to 0 as x goes to 0, so that the following holds. For all $\varepsilon > 0$, and any instance of the problem with a risk assignment satisfying the ε-approximate versions of fairness conditions (A), (B), and (C), the instance must satisfy either the $f(\varepsilon)$-approximate version of perfect prediction or the $f(\varepsilon)$-approximate version of equal base rates.*

Thus, anything that approximately satisfies the fairness constraints must approximately look like one of the two simple cases identified above.

Finally, in connection to Theorem 1.1, we note that when the two groups have equal base rates, then one can ask for the most accurate risk assignment that satisfies all three fairness conditions (A), (B), and (C) simultaneously. Since the risk assignment that gives the same score to everyone satisfies the three conditions, we know that at least one such risk assignment exists; hence, it is natural to seek to optimize over the set of all such assignments. We consider this algorithmic question in the final technical section of the paper.

To reflect a bit further on our main theorems and what they suggest, we note that our intention in the present work isn't to make a recommendation on how conflicts between different definitions of fairness should be handled. Nor is our intention to analyze which definitions of fairness are violated in particular applications or datasets. Rather, our point is to establish certain unavoidable trade-offs between the definitions, regardless of the specific context and regardless of the method used to compute risk scores. Since each of the definitions reflect (and have been proposed as) natural notions of what it should mean for a risk score to be fair, these trade-offs suggest a striking implication: that outside of narrowly delineated cases, any assignment of risk scores can in principle be subject to natural criticisms on the grounds of bias. This is equally true whether the risk score is determined by an algorithm or by a system of human decision-makers.

Special Cases of the Model. Our main results, which place strong restrictions on when the three fairness conditions can be simultaneously satisfied, have more power when the underlying model of the input is more general, since it means that the restrictions implied by the theorems apply in greater generality. However, it is also useful to note certain special cases of our model, obtained by limiting the flexibility of certain parameters in intuitive ways. The point is that our results apply *a fortiori* to these more limited special cases.

First, we have already observed one natural special case of our model: cases in which, for each feature vector σ, only members of one group (but not the other) can exhibit σ. This means that σ contains perfect information about group membership, and so it corresponds to instances in which risk assignments would have the potential to use knowledge of an individual's group membership. Note that we can convert any instance of our problem into a new instance that belongs to this special case as follows. For each feature vector σ, we create two new feature vectors $\sigma^{(1)}$ and $\sigma^{(2)}$; then, for each member of group 1 who had feature vector σ, we assign them $\sigma^{(1)}$, and for each member of group 2 who had feature vector σ, we assign them

$\sigma^{(2)}$. The resulting instance has the property that each feature vector is associated with members of only one group, but it preserves the essential aspects of the original instance in other respects.

Second, we allow risk assignments in our model to split people with a given feature vector σ over several bins. Our results also therefore apply to the natural special case of the model with *integral* risk assignments, in which all people with a given feature σ must go to the same bin.

Third, our model is a generalization of binary classification, which only allows for 2 bins. Note that although binary classification does not explicitly assign scores, we can consider the probability that an individual belongs to the positive class given that they were assigned to a specific bin to be the score for that bin. Thus, our results hold in the traditional binary classification setting as well.

Data-Generating Processes. Finally, there is the question of where the data in an instance of our problem comes from. Our results do not assume any particular process for generating the positive/negative class labels, feature vectors, and group memberships; we simply assume that we are given such a collection of values (regardless of where they came from), and then our results address the existence or non-existence of certain risk assignments for these values.

This increases the generality of our results, since it means that they apply to any process that produces data of the form described by our model. To give an example of a natural generative model that would produce instances with the structure that we need, one could assume that each individual starts with a "hidden" class label (positive or negative), and a feature vector σ is then probabilistically generated for this individual from a distribution that can depend on their class label and their group membership. (If feature vectors produced for the two groups are disjoint from one another, then the requirement that the value of p_σ is independent of group membership given σ necessarily holds.) Since a process with this structure produces instances from our model, our results apply to data that arises from such a generative process.

It is also interesting to note that the basic set-up of our model, with the population divided across a set of feature vectors for which race provides no additional information, is in fact a very close match to the information one gets from the output of a well-calibrated risk tool. In this sense, one setting for our model would be the problem of applying post-processing to the output of such a risk tool to ensure additional fairness guarantees. Indeed, since much of the recent controversy about fair risk scores has involved risk tools that are well-calibrated but lack the other fairness conditions we consider, such an interpretation of the model could be a useful way to think about how one might work with these tools in the context of a broader system.

1.3 Further Related Work

Mounting concern over discrimination in machine learning has led to a large body of new work seeking to better understand and prevent it. Barocas and Selbst survey a range of ways in which data-analysis algorithms can lead to discriminatory outcomes [3], and review articles by Romei and Ruggieri [25] and Zliobaite [30] survey data-analytic and algorithmic methods for measuring discrimination.

Kamiran and Calders [21] and Hajian and Domingo-Ferrer [18] seek to modify datasets to remove any information that might permit discrimination. Similarly, Zemel et al. look to learn fair intermediate representations of data while preserving information needed for classification [29]. Joseph et al. consider how fairness issues can arise during the process of learning, modeling this using a multi-armed bandit framework [20].

Fair prediction with disparate impact:
A study of bias in recidivism prediction instruments

Alexandra Chouldechova *

Last revised: February 8, 2017

Abstract

Recidivism prediction instruments (RPI's) provide decision makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time. While such instruments are gaining increasing popularity across the country, their use is attracting tremendous controversy. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. This paper discusses several fairness criteria that have recently been applied to assess the fairness of recidivism prediction instruments. We demonstrate that the criteria cannot all be simultaneously satisfied when recidivism prevalence differs across groups. We then show how disparate impact can arise when a recidivism prediction instrument fails to satisfy the criterion of error rate balance.

Keywords: disparate impact; bias; recidivism prediction; risk assessment; fair machine learning

1 Introduction

Risk assessment instruments are gaining increasing popularity within the criminal justice system, with versions of such instruments being used or considered for use in pre-trial decision-making, parole decisions, and in some states even sentencing[1,2,3]. In each of these cases, a high-risk classification—particularly a high-risk misclassification—may have a direct adverse impact on a criminal defendant's outcome. If the use of RPI's is to become commonplace, it is especially important to ensure that the instruments are free from discriminatory biases that could result in unethical practices and inequitable outcomes for different groups.

In a recent widely popularized investigation conducted by a team at ProPublica, Angwin et al.[4] studied an RPI called COMPAS[a], concluding that it is biased against black defendants. The authors

*Heinz College, Carnegie Mellon University

[a]COMPAS[5] is a risk assessment instrument developed by Northpointe Inc.. Of the 22 scales that COMPAS provides, the Recidivism risk and Violent Recidivism risk scales are the most widely used. The empirical results in this paper are based on decile scores coming from the COMPAS Recidivism risk scale.

found that the likelihood of a non-recidivating black defendant being assessed as high risk is nearly twice that of white defendants. Similarly, the likelihood of a recidivating black defendant being assessed as low risk is nearly half that of white defendants. In technical terms, these findings indicate that the COMPAS instrument has considerably higher false positive rates and lower false negative rates for black defendants than for white defendants.

ProPublica's analysis has met with much criticism from both the academic community and from the Northpointe corporation. Much of the criticism has focussed on the particular choice of fairness criteria selected for the investigation. Flores et al.[6] argue that the correct approach for assessing RPI bias is instead to check for *calibration*, a fairness criterion that they show COMPAS satisfies. Northpointe in their response[7] argue for a still different approach that checks for a fairness criterion termed *predictive parity*, which they demonstrate COMPAS also satisfies. We provide precise definitions and a more in-depth discussion of these and other fairness criteria in Section 2.1.

In this paper we show that the differences in false positive and false negative rates cited as evidence of racial bias by Angwin et al.[4] are a direct consequence of applying an RPI that that satisfies predictive parity to a population in which recidivism prevalence[a] differs across groups. Our main contribution is twofold. (1) First, we make precise the connection between the predictive parity criterion and error rates in classification. (2) Next, we demonstrate how using an RPI that has different false postive and false negative rates between groups can lead to disparate impact when individuals assessed as high risk receive stricter penalties. Throughout our discussion we use the term *disparate impact* to refer to settings where a penalty policy has unintended disproportionate adverse impact on a particular group.

It is important to bear in mind that fairness itself—along with the notion of disparate impact—is a social and ethical concept, not a statistical one. A risk prediction instrument that is fair with respect to particular fairness criteria may nevertheless result in disparate impact depending on how and where it is used. In this paper we consider hypothetical use cases in which we are able to directly connect particular fairness properties of an RPI to a measure of disparate impact. We present both theoretical and empirical results to illustrate how disparate impact can arise.

1.1 Outline of paper

We begin in Section 2 by providing some background on several of the different fairness criteria that have appeared in recent literature. We then proceed to demonstrate that an instrument that satisfies predictive parity cannot have equal false positive and negative rates across groups when the recidivism prevalence differs across those groups. In Section 3 we analyse a simple risk assessment-based sentencing policy and show how differences in false positive and false negative rates can result in disparate impact under this policy. In Section 3.3 we back up our theoretical analysis by presenting some empirical results based on the data made available by the ProPublica investigators. We conclude with a discussion of the issues that biased data presents for the arguments put forth in this paper.

---

[a] *Prevalence*, also termed the *base rate*, is the proportion of individuals who recidivate in a given population.

1.2 Data description and setup

The empirical results in this paper are based on the Broward County data made publicly available by ProPublica[8]. This data set contains COMPAS recidivism risk decile scores, 2-year recidivism outcomes, and a number of demographic and crime-related variables on individuals who were scored in 2013 and 2014. We restrict our attention to the subset of defendants whose race is recorded as African-American (b) or Caucasian (w).[a] After applying the same data pre-processing and filtering as reported in the ProPublica analysis, we are left with a data set on $n = 6150$ individuals, of whom $n_b = 3696$ are African-American and $n_c = 2454$ are Caucasian.

2 Assessing fairness

2.1 Background

We begin by with some notation. Let $S = S(x)$ denote the risk score based on covariates $X = x \in \mathbb{R}^p$, with higher values of S corresponding to higher levels of assessed risk. We will interchangeably refer to S as a *score* or an *instrument*. For simplicity, our discussion of fairness criteria will focus on a setting where there exist just two groups. We let $R \in \{b, w\}$ denote the group to which an individual belongs, and do not preclude R from being one of the elements of X. We denote the outcome indicator by $Y \in \{0, 1\}$, with $Y = 1$ indicating that the given individual goes on to recidivate. Lastly, we introduce the quantity s_{HR}, which denotes the high-risk score threshold. Defendants whose score S exceeds s_{HR} will be referred to as *high-risk*, while the remaining defendants will be referred to as *low-risk*.

With this notation in hand, we now proceed to define and discuss several fairness criteria that commonly appear in the literature, beginning with those mentioned in the introduction. We indicate cases where a given criterion is known to us to also commonly appear under some other name. All of the criteria presented below can also be assessed *conditionally* by further conditioning on some covariates in X. We discuss this point in greater detail in Section 3.1.

Definition 1 (Calibration)**.** A score $S = S(x)$ is said to be *well-calibrated* if it reflects the same likelihood of recidivism irrespective of the individuals' group membership. That is, if for all values of s,

$$\mathbb{P}(Y = 1 \mid S = s, R = b) = \mathbb{P}(Y = 1 \mid S = s, R = w). \tag{2.1}$$

Within the educational and psychological testing and assessment literature, the notion of *calibration* features among the widely accepted and adopted standards for empirical fairness assessment. In this literature, an instrument that is *well-calibrated* is referred to as being *free from predictive bias*. This criterion has recently been applied to the PCRA[b] instrument, with initial findings suggesting that calibration is satisfied with respect race[10,11], but not with respect to gender[12]. In

[a] There are 6 racial groups represented in the data. 85% of individuals are either African-American or Caucasian.

[b] The Post Conviction Risk Assessment (PCRA) tool was developed by the Administrative Office of the United States Courts for the purpose of improving "the effectiveness and efficiency of post-conviction supervision" [9]

their response to the ProPublica investigation, Flores et al.[6] verify that COMPAS is well-calibrated using logistic regression modeling.

Definition 2 (Predictive parity)**.** A score $S = S(x)$ satisfies *predictive parity* at a threshold s_{HR} if the likelihood of recidivism among high-risk offenders is the same regardless of group membership. That is, if,

$$\mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = b) = \mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = w). \tag{2.2}$$

Predictive parity at a given threshold s_{HR} amounts to requiring that the *positive predictive value* (PPV) of the classifier $\hat{Y} = \mathbb{1}_{S > s_{\text{HR}}}$ be the same across groups. While predictive parity and calibration look like very similar criteria, well-calibrated scores can fail to satisfy predictive parity at a given threshold. This is because the relationship between (2.2) and (2.1) depends on the conditional distribution of $S \mid R = r$, which can differ across groups in ways that result in PPV imbalance. In the simple case where S itself is binary, a score that is well-calibrated will also satisfy predictive parity. Northpointe's refutation[7] of the ProPublica analysis shows that COMPAS satisfies predictive parity for threshold choices of interest.

Definition 3 (Error rate balance)**.** A score $S = S(x)$ satisfies *error rate balance* at a threshold s_{HR} if the false positive and false negative error rates are equal across groups. That is, if,

$$\mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = b) = \mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = w)\ , \quad \text{and} \tag{2.3}$$
$$\mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = b) = \mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = w), \tag{2.4}$$

where the expressions in the first line are the group-specific false positive rates, and those in the second line are the group-specific false negative rates.

ProPublica's analysis considered a threshold of $s_{\text{HR}} = 4$, which they showed leads to considerable imbalance in both false positive and false negative rates. While this choice of cutoff met with some criticism, we will see later in this section that error rate imbalance persists—indeed, must persist—for any choice of cutoff at which the score satisfies the predictive parity criterion. Error rate balance is also closely connected to the notions of *equalized odds* and *equal opportunity* as introduced in the recent work of Hardt et al.[13].

Definition 4 (Statistical parity)**.** A score $S = S(x)$ satisfies *statistical parity* at a threshold s_{HR} if the proportion of individuals classified as high-risk is the same for each group. That is, if,

$$\mathbb{P}(S > s_{\text{HR}} \mid R = b) = \mathbb{P}(S > s_{\text{HR}} \mid R = w) \tag{2.5}$$

Statistical parity also goes by the name of *equal acceptance rates*[14] or *group fairness*[15], though it should be noted that these terms are in many cases not used synonymously. While our discussion focusses primarily on first three fairness criteria, statistical parity is widely used within the machine learning community and may be the criterion with which many readers are most familiar[16,17]. Statistical parity is well-suited to contexts such as employment or admissions, where it may be desirable or required by law or regulation to employ or admit individuals in equal proportion across racial, gender, or geographical groups. It is, however, a difficult criterion to motivate in the recidivism prediction setting, and thus will not be further considered in this work.

2.2 Further related work

Though the study of discrimination in decision making and predictive modeling is rapidly evolving, it also has a long and rich multidisciplinary history. Romei and Ruggieri [18] provide an excellent overview of some of the work in this broad subject area. The recent work of Barocas and Selbst [19] offers a broad examination of algorithmic fairness framed within the context of anti-discrimination laws governing employment practices. Hannah-Moffat [20], Skeem [21], and Monahan and Skeem [22] examine legal and ethical issues relating specifically to the use of risk assessment instruments in sentencing, citing the potential for race and gender discrimination as a major concern.

In work concurrent with our own, several other researchers have also investigated the compatibility of different notions of fairness. Kleinberg et al. [23] show that calibration cannot be satisfied simultaneously with the fairness criteria of *balance for the negative class* and *balance for the positive class*. Translated into the present context, the latter criteria require that the average score assigned to non-recidivists (the negative class) should be the same for both groups, and that the same should hold among recidivists (the positive class). The work of Corbett-Davies et al. [24] closely parallels the results that we present in Section 2.3, reaching the same conclusion regarding the incompatibility of predictive parity and error rate balance in the setting of unequal prevalence.

2.3 Predictive parity, false positive rates, and false negative rates

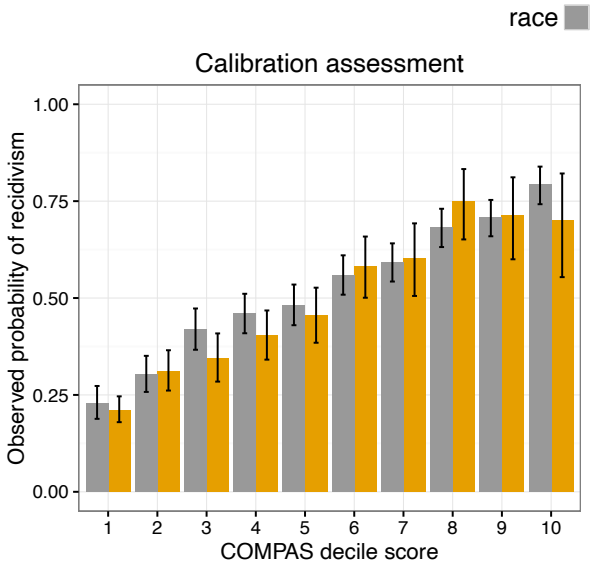
In this section we present our first main result, which establishes that predictive parity is incompatible with error rate balance when prevalence differs across groups. To better motivate the discussion, we begin by presenting an empirical fairness assessment of the COMPAS RPI. Figure 1 shows plots of the observed recidivism rates and error rates corresponding to the fairness notions of calibration, predictive parity, and error rate balance. We see that the COMPAS RPI is (approximately) well-calibrated, and also satisfies predictive parity provided that the high-risk cutoff s_{HR} is 4 or greater. However, COMPAS fails on both false positive and false negative error rate balance across the range of high-risk cutoffs.

Angwin et al. [4] focussed on a high-risk cutoff of $s_{\text{HR}} = 4$ for their analysis, which some critics have argued is too low, suggesting that $s_{\text{HR}} = 7$ is more suitable. As can be seen from Figures 1c and 1d, significant error rate imbalance persists at this cut-off as well. Moreover, the error rates achieved at so high a cutoff are at odds with evidence suggesting that the use of RPI's is of interest in settings where false negatives have a higher cost than false positives, with relative cost estimates ranging from 2.6 to upwards of 15. [25,26]

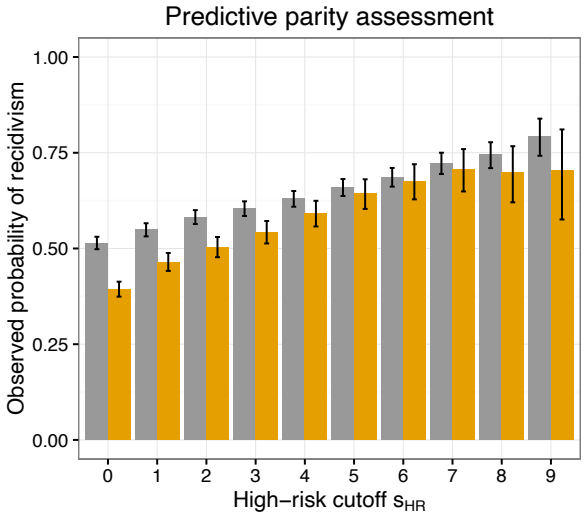
As we now proceed to show, the error rate imbalance exhibited by COMPAS is not a coincidence, nor can it be remedied in the present context. When the recidivism prevalence–i.e., the base rate $\mathbb{P}(Y = 1 \mid R = r)$—differs across groups, any instrument that satisfies predictive parity at a given threshold s_{HR} *must* have imbalanced false positive or false negative errors rates at that threshold. To understand why predictive parity and error rate balance are mutually exclusive in the setting of unequal recidivism prevalence, it is instructive to think of how these quantities are all related.

Given a particular choice of s_{HR}, we can summarize an instrument's performance in terms of a confusion matrix, as shown in Table 1 below.
All of the fairness metrics presented in Section 2.1 can be thought of as imposing constraints on

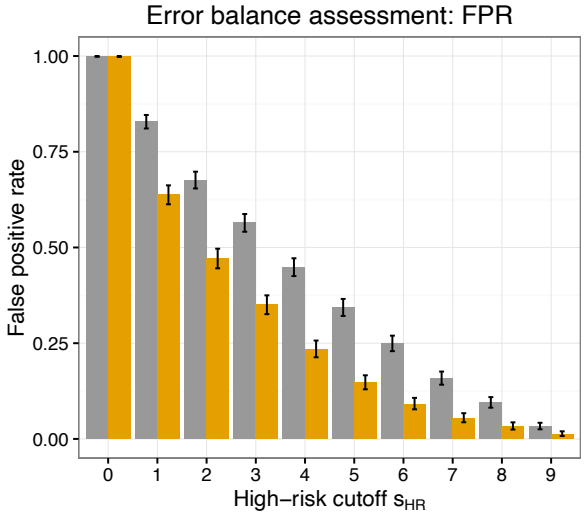
the values (or the distribution of values) in this table. Another constraint—one that we have no direct control over—is imposed by the recidivism prevalence within groups. It is not difficult to



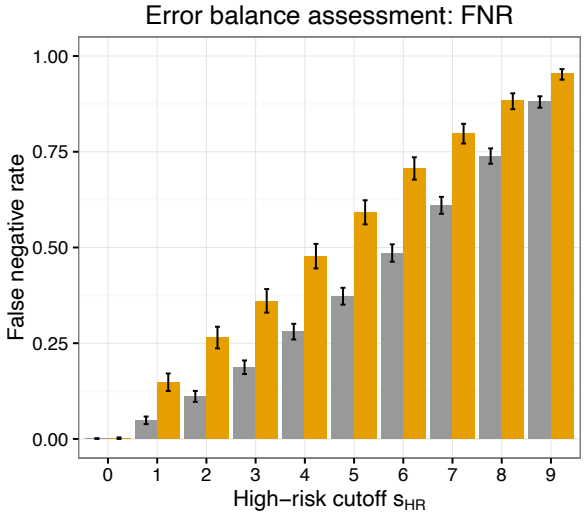
(a) Bars represent empirical estimates of the expressions in (2.1): $\mathbb{P}(Y = 1 \mid S = s, R = r)$ for decile scores $s \in \{1, \ldots, 10\}$.



(b) Bars represent empirical estimates of the expressions in (2.2): $\mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = r)$ for values of the high-risk cutoff $s_{\text{HR}} \in \{0, \ldots, 9\}$



(c) Bars represent observed false positive rates, which are empirical estimates of the expressions in (2.3): $\mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = r)$ for values of the high-risk cutoff $s_{\text{HR}} \in \{0, \ldots, 9\}$



(d) Bars represent observed false negative rates, which are empirical estimates of the expressions in (2.4): $\mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = r)$ for values of the high-risk cutoff $s_{\text{HR}} \in \{0, \ldots, 9\}$

Figure 1: Empirical assessment of the COMPAS RPI according to three of the fairness criteria presented in Section 2.1. Error bars represent 95% confidence intervals. These Figures confirm that COMPAS is (approximately) well-calibrated, satisfies predictive parity for high-risk cutoff values of 4 or higher, but fails to have error rate balance.

	Low-Risk	High-Risk
$Y = 0$	TN	FP
$Y = 1$	FN	TP

Table 1: T/F denote True/False and N/P denote Negative/Positive. For instance, FP is the number of false positives: individuals who are classified as high-risk but who do not reoffend.

show that the prevalence (p), positive predictive value (PPV), and false positive and negative error rates (FPR, FNR) are related via the equation

$$\text{FPR} = \frac{p}{1-p}\frac{1-\text{PPV}}{\text{PPV}}(1-\text{FNR}). \tag{2.6}$$

From this simple expression we can see that if an instrument satisfies predictive parity—that is, if the PPV is the same across groups—but the prevalence differs between groups, the instrument cannot achieve equal false positive and false negative rates across those groups.

This observation enables us to better understand why we observe such large discrepancies in FPR and FNR between black and white defendants in Figure 1. The recidivism rate among black defendants in the data is 51%, compared to 39% for White defendants. Thus at any threshold s_{HR} where the COMPAS RPI satisfies predictive parity, equation (2.6) tells us that some level of imbalance in the error rates must exist. Since not all of the fairness criteria can be satisfied at the same time, it becomes important to understand the potential impact of failing to satisfy particular criteria. This question is explored in the context of a hypothetical risk-based sentencing framework in the next section.

3 Assessing impact

In this section we show how differences in false positive and false negative rates can result in disparate impact under policies where a high-risk assessment results in a stricter penalty for the defendant. Such situations may arise when risk assessments are used to inform bail, parole, or sentencing decisions. In Pennsylvania and Virginia, for instance, statutes permit the use of RPI's in sentencing, provided that the sentence ultimately falls within accepted guidelines[1]. We use the term "penalty" somewhat loosely in this discussion to refer to outcomes both in the pre-trial and post-conviction phase of legal proceedings. For instance, even though pre-trial outcomes such as the amount at which bail is set are not punitive in a legal sense, we nevertheless refer to bail amount as a "penalty" for the purpose of our discussion.

There are notable cases where RPI's are used for the express purpose of informing risk reduction efforts. In such settings, individuals assessed as high risk receive what may be viewed as a benefit rather than a penalty. The PCRA score, for instance, is intended to support precisely this type of decision-making at the federal courts level[11]. Our analysis in this section specifically addresses use cases where high-risk individuals receive stricter penalties.

To begin, consider a setting in which guidelines indicate that a defendant is to receive a penalty