Interactive Visualization of Fairness Tradeoffs

Jonathan Stray
Partnership on AI
jonathan@partnershiponai.com

Karen Hao
MIT Tech Review
Karen.Hao@technologyreview.com

ABSTRACT

ProPublica's groundbreaking Machine Bias investigation [1] showed that the COMPAS algorithm was incorrectly predicting re-arrest for black defendants at twice the rate of white defendants. Subsequent work on fairness in machine learning showed that, if the re-arrest rates in the training data differ between races, any prediction algorithm which assigns the same risk score to people who are, in fact, equally likely to be re-arrested will necessarily produce differing false positive rates [2][3].

This dry technical conclusion has profound real-world implications: it is simply not possible to build an algorithm that is racially fair according to all measures. We set out to explain this fact, and explain *why* this is true, in the hope of advancing the public conversation around the use of these tools. We began by reviewing the legal and statistical literature on this topic, then interviewed key researchers such as Mayson who has a written detailed explanation of these tradeoffs [4] and Stevenson who has studied risk assessment in the real world [5].

Our first whiteboard sketch was derived from an illustration in Mayson's paper (figure 1). We translated this into a prototype interactive using Observable, but user testing revealed that even our tech-savvy colleagues could not understand it.

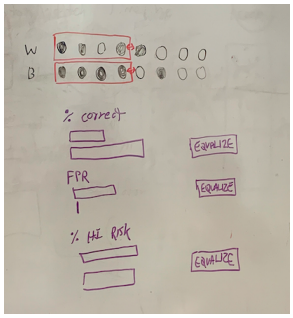

Figure 1: Our initial whiteboard design

We went through several intermediate designs and eventually found an approach that users could understand. Even so, the concept was complicated enough that we chose to break the piece into a series of interactive steps, starting with the idea of grouping people by their risk score, then making decisions based on a threshold (figure 2), introducing different kinds of fairness measures, splitting the results by race, and finally comparing multiple fairness measures across multiple races (figure 3).

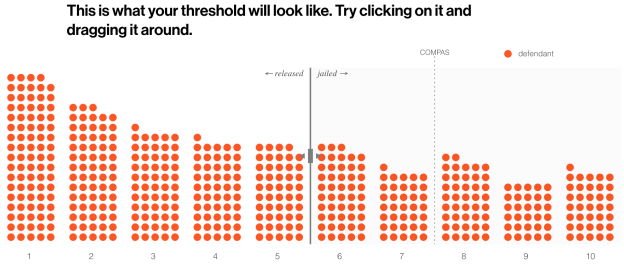

Figure 2: An early step of the final sequence of interactives

The final piece [6] is a hybrid text and interactive narrative which builds up slowly to a sophisticated visual explanation of the inherent tradeoffs of prediction. And it even works on mobile.

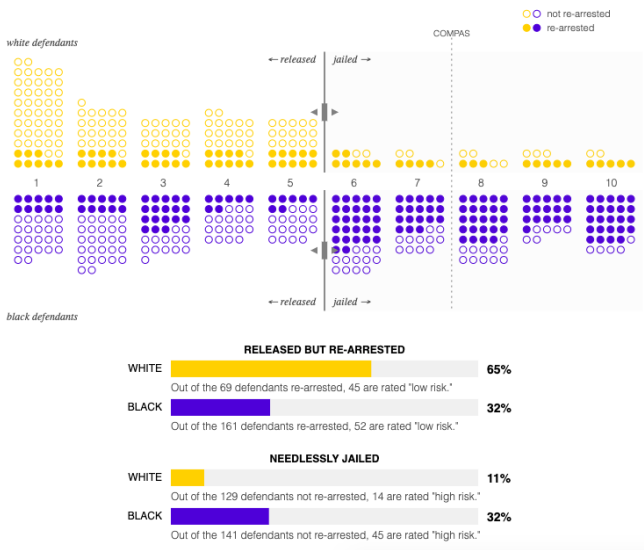

Figure 3: The final interactive step

REFERENCES

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," *ProPublica*, 2016.
[2] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," 2017.
[3] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in Criminal Justice Risk Assessments: The State of the Art," pp. 1–42, 2017.
[4] S. G. Mayson, "Bias In, Bias Out," *Yale Law J.*, 2019.
[5] M. Stevenson, "Assessing Risk Asessment in Action," 2017.
[6] K. Hao and J. Stray, "Can you make AI fairer than a judge? Play our courtroom algorithm game - MIT Technology Review," *MIT Technology Review*, 2019.