Blind Justice:

Algorithmically Masking Race in Charging Decisions

Alex Chohlas-Wood, Joe Nudell, Keniel Yao, Zhiyuan "Jerry" Lin, Julian Nyarko, Sharad Goel
Stanford University Computational Policy Lab

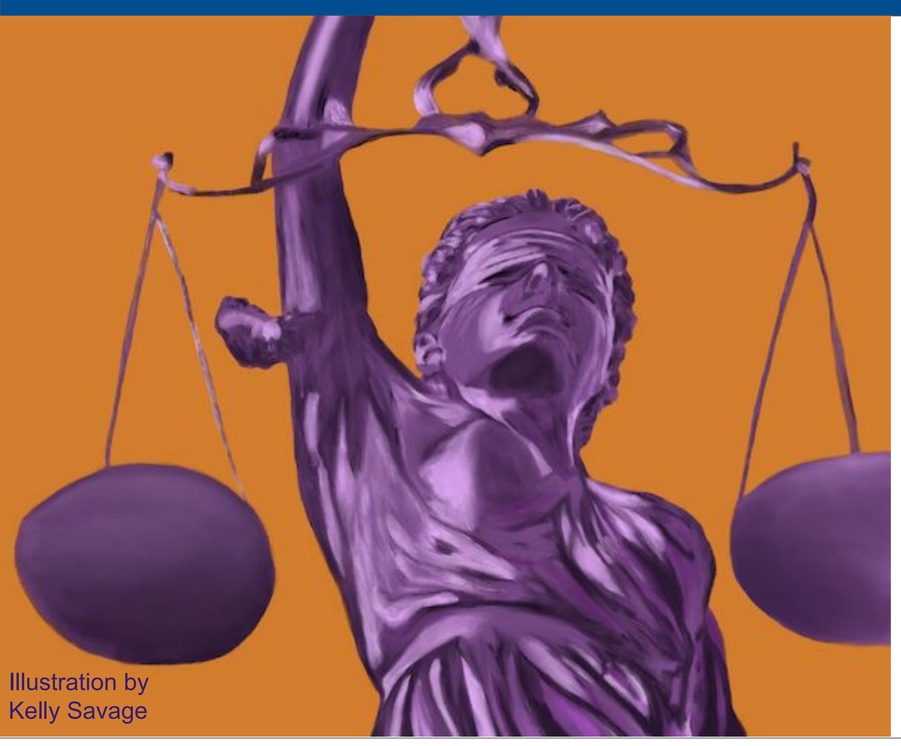


Illustration by Kelly Savage

Summary

- We designed an algorithm that automatically redacts race-related information from police incident narratives, and deployed the algorithm at a district attorney's office in a major American city.
- We demonstrated that our redaction algorithm makes it difficult for an expert annotator to infer a suspect's race from the redacted narrative.
- However, charging decisions at our partner DA's office did not show evidence of bias before our pilot. So—perhaps unsurprisingly—we do not yet know whether blind charging affects charging rates by race.

A toy example

Lucy Johnson reported that a **Black** male with **brown hair** wearing a black jacket assaulted her in **Midtown**, next to **Johnson's** home. She reported the incident to **Officer Lee**.

Original narrative

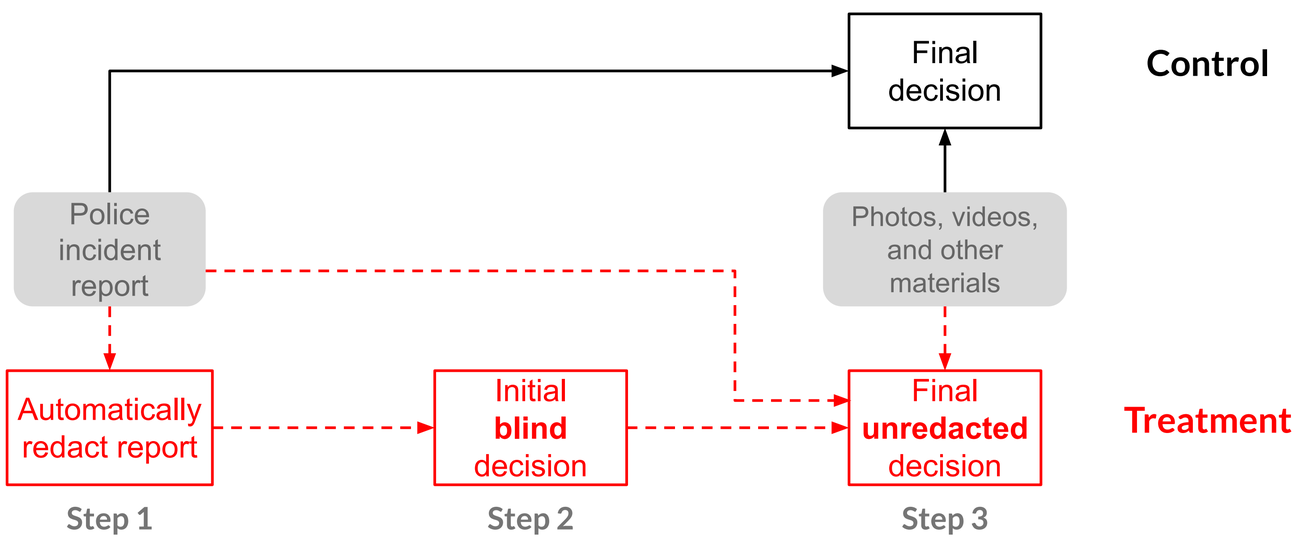
→

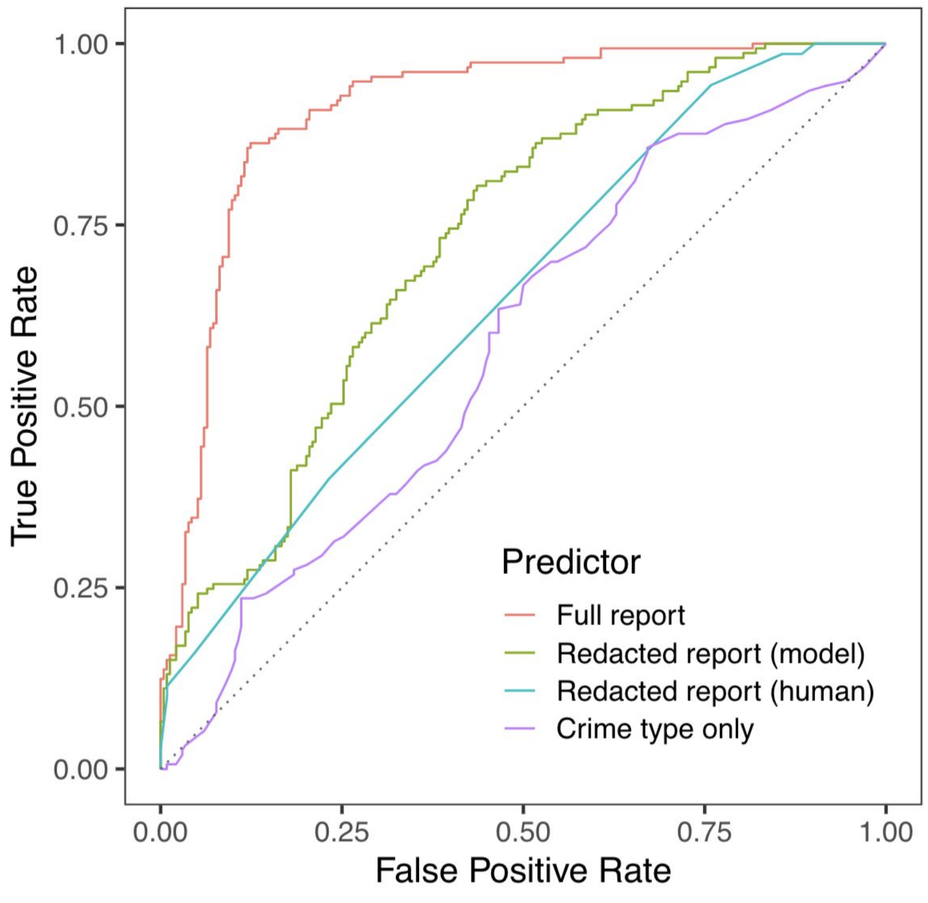
[Victim 1] reported that a **[race]** male with **[hair color]** wearing a black jacket assaulted her in **[neighborhood]**, next to **[Victim 1]'s** home. She reported the incident to **[Officer 1]**.

Automatically redacted narrative

Experimental setup

- In our treatment arm, prosecutors reviewed cases twice: first, they conducted a preliminary race-obscured review with the redacted incident report; and later, they engaged in a final review with all available (unredacted) information.
- For our analysis, we compared differences in final decisions between the treatment and control arms.





Assessing redaction quality

- We evaluated whether it's possible to predict the race of the suspect involved in each incident.
- Better performance is indicated by curves that approach the upper left corner; the dotted diagonal line indicates a baseline model that guesses race completely at random.
- Both the human annotator and model with access to redacted information perform comparably to the baseline crime type model, while the model with access to all information performs substantially better than both.
- This suggests our algorithm is able to effectively redact race-related information from incident reports.