



TALLER
¿Qué entiendes
por Justicia?

ANTECEDENTES (I)

- Hace más de una década que en los juzgados y tribunales de EEUU se utilizan algoritmos de machine learning para predecir la probabilidad de reincidencia de las personas detenidas.
- Con los resultados en la mano, jueces de todo EEUU deciden si los acusados pueden salir o no en libertad condicional y las cantidades de la fianza que deben aportar para ello.
- El uso de estos algoritmos se ha extendido por todo EEUU sin previamente testear de forma rigurosa su rendimiento ni analizar su impacto sobre los diferentes grupos de población a los que se aplica.
- Hay dos herramientas comerciales líderes a nivel nacional , una de las cuales es el software **COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)** desarrollado y comercializado por Northpointe, Inc., una compañía con ánimo de lucro.

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)

- El modelo arroja un conjunto de puntuaciones que van del 1 (mínimo) a 10 (máximo) riesgo de reincidencia a partir de las respuestas a un cuestionario de 137 preguntas, facilitadas por los acusados o extraídas de los antecedentes penales.
- Como exige la ley aplicable, la raza de detenido NO es una de las preguntas del cuestionario.
- La encuesta pregunta a los acusados cosas como: "¿Alguna vez uno de sus padres fue enviado a la cárcel o prisión?" "Dispones de un domicilio habitual?" "¿Cuántos de tus amigos / conocidos están consumiendo drogas ilegalmente?" y "¿Con qué frecuencia te peleaste en la escuela?" El cuestionario también pide a las personas que estén de acuerdo o en desacuerdo con afirmaciones como "Una persona hambrienta tiene derecho a robar" y "Si la gente me hace enojar o perder los estribos, puedo ser peligroso".

Ejemplo de cuestionario:

<https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE>

- Las respuestas al cuestionario se introducen en el software COMPAS para generar varios puntuaciones que incluyen predicciones de "Riesgo de reincidencia" y "Riesgo de reincidencia violenta".

The image shows a sample COMPAS Risk Assessment form. It is a structured questionnaire with multiple sections. The top section is titled 'Risk Assessment' and contains fields for 'Name', 'Date of Birth', 'Date of Arrest', 'Date of Release', 'Date of Assessment', and 'Assessor'. Below this, there are several sections of questions, each with a 'Yes', 'No', or 'Don't Know' response option. The questions cover various aspects of the offender's background, including family history, employment, substance use, and criminal history. The form is partially filled out with blacked-out text, indicating that it is a sample form.

ANTECEDENTES (II)

- La empresa no divulga los cálculos utilizados para llegar a las puntuaciones de riesgo de los acusados, es decir, **el modelo de machine learning no es público**. Se considera secreto comercial. Por tanto, el funcionamiento del sistema es opaco para los acusados y para la sociedad en general.
- En mayo de 2016, ProPublica, una organización sin ánimo de lucro y ganadora de un Premio Pulitzer, publica un artículo analizando el impacto de COMPAS desde el punto de vista racial, en particular el impacto sobre las personas de raza negra.
- Este es primer análisis que se hizo de este algoritmo desde el punto de vista de equidad (Fairness).

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

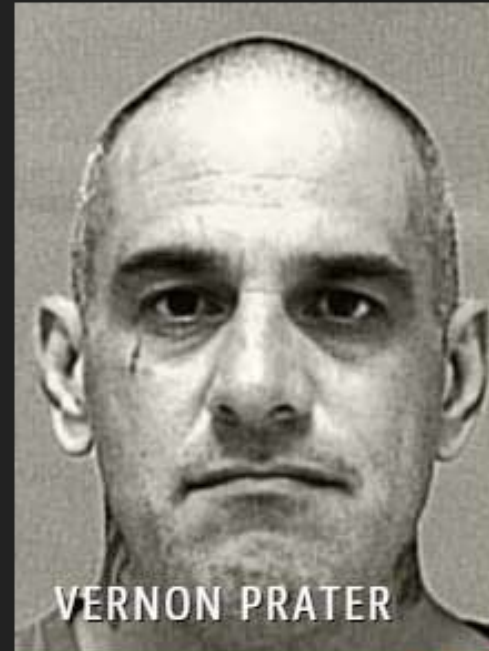
May 23, 2016

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Methodology:

<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

Two Petty Theft Arrests



VERNON PRATER

LOW RISK

3



BRISHA BORDEN

HIGH RISK

8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

- A través de una solicitud de registros públicos, ProPublica obtuvo dos años de puntuaciones COMPAS de la Oficina del Sheriff del Condado de Broward en Florida. Recibieron datos de las 18,610 personas que fueron calificadas en 2013 y 2014.
- Tras un proceso de limpieza, el número de personas en la muestra se redujo a 11,757 personas.
- Compararon las puntuaciones de riesgo de reincidencia arrojadas por la herramienta COMPAS con las tasas reales de reincidencia de los acusados en los dos años posteriores a su calificación.
- Descubrieron que la herramienta predijo correctamente la reincidencia en los dos años siguientes en el 61% de los casos, y solo fue correcta en sus predicciones de reincidencia violenta el 20% de los casos.



Accuracy:

Del riesgo de reincidencia : 61%

Del riesgo de reincidencia violenta: 20%

All Defendants			Black Defendants			White Defendants		
	Low	High		Low	High		Low	High
Survived	2681	1282	Survived	990	805	Survived	1139	349
Recidivated	1216	2035	Recidivated	532	1369	Recidivated	461	505
FP rate: 32.35			FP rate: 44.85			FP rate: 23.45		
FN rate: 37.40			FN rate: 27.99			FN rate: 47.72		
PPV: 0.61			PPV: 0.63			PPV: 0.59		
NPV: 0.69			NPV: 0.65			NPV: 0.71		
LR+: 1.94			LR+: 1.61			LR+: 2.23		
LR-: 0.55			LR-: 0.51			LR-: 0.62		

Riesgo de reincidencia, ProPublica

Si nos fijamos en los **ACIERTOS** del modelo (True Positives y True Negatives), el rendimiento de COMPAS es similar para los dos grupos.

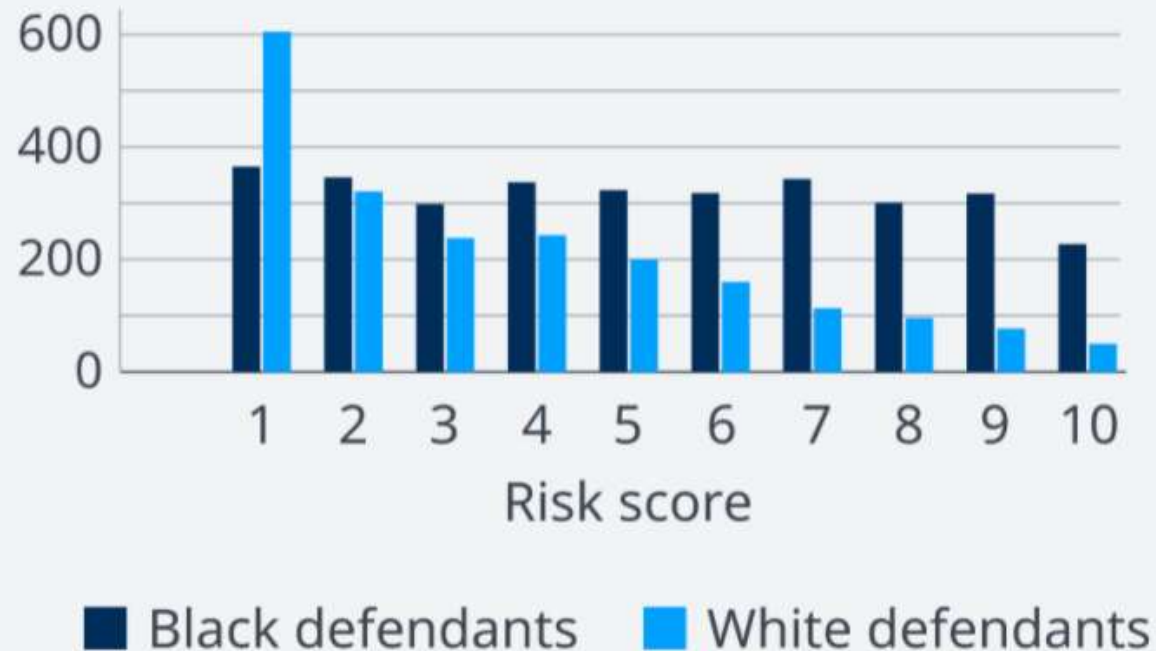
La probabilidad de recibir una predicción positiva correcta (Positive Predictive Value o PPV) y la probabilidad de recibir una predicción negativa correcta (Negative predicted value o NPV) no difiere mucho en ambos grupos.

[https://en.wikipedia.org/wiki/Fairness_\(machine_learning\)](https://en.wikipedia.org/wiki/Fairness_(machine_learning))

El problema está en los errores...

Risk scores: black vs white defendants

Number of defendants in each risk-score group



Source: ProPublica | data from Broward County, Fla.

©DW

Distribución de las puntuaciones para el riesgo de reincidencia de 6.172 detenidos que en realidad NO reincidieron en el plazo de dos años.

ProPublica
Machine Bias
Marzo 2016

All Defendants			Black Defendants			White Defendants		
	Low	High		Low	High		Low	High
Survived	2681	1282	Survived	990	805	Survived	1139	349
Recidivated	1216	2035	Recidivated	532	1369	Recidivated	461	505
FP rate: 32.35			FP rate: 44.85			FP rate: 23.45		
FN rate: 37.40			FN rate: 27.99			FN rate: 47.72		
PPV: 0.61			PPV: 0.63			PPV: 0.59		
NPV: 0.69			NPV: 0.65			NPV: 0.71		
LR+: 1.94			LR+: 1.61			LR+: 2.23		
LR-: 0.55			LR-: 0.51			LR-: 0.62		

Riesgo de reincidencia, ProPublica

Sin embargo, si nos fijamos en los **ERRORES** del modelo (False Positives y False Negatives), el rendimiento del modelo es muy diferente para cada grupo.

- Los acusados negros que no reinciden tenían casi el doble de probabilidades de ser clasificados por COMPAS como de mayor riesgo en comparación con sus homólogos blancos (False Positive Rate o FPR de un 45% frente a 23%).
- COMPAS clasificó erroneamente a los reincidentes blancos como de bajo riesgo un 70,5 % más a menudo que a los reincidentes negros (False negative rate o FNR de un 48% frente al 28%).

Análisis

1. ¿Qué tipo o tipos de daños consideráis que está produciendo el modelo?
2. La raza no se utilizó como variable a los efectos de entrenar o testear el algoritmo. ¿Cuáles pensáis que fueron las proxies?
3. ¿Qué sesgos podéis identificar? ¿en qué fases de desarrollo del modelo pensáis que se están introduciendo los sesgos?
4. ¿Qué sesgos creéis que pueden estar distorsionando la aplicación del modelo por los jueces?
5. ¿Qué os parece la accuracy del modelo?
6. ¿Qué otros problemas detectáis en la forma en que se ha implementado el modelo en el sistema penal norteamericano?



Pero...es más
complicado de lo que
parece...

Refutaciones

- Nortpoint.Inc research department, July 8, 2016.

“In its paper, Northpointe dismissed the racial disparities we detected by saying “this pattern does not show evidence of bias, but rather is a natural consequence of using unbiased scoring rules for groups that happen to have different distributions of scores.”

Es decir, que la tasa de reincidencia **para personas negras y blancos en los EEUU son diferentes, así que la distribución de los errores es necesariamente diferente.**

http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

- *“False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And it’s Biased Against Blacks.”*

Anthony W. Flores, Ph.D, Christopher T. Lowenkamp, Ph.D., Kristin Bechtel, M.S.

http://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf

Defensa

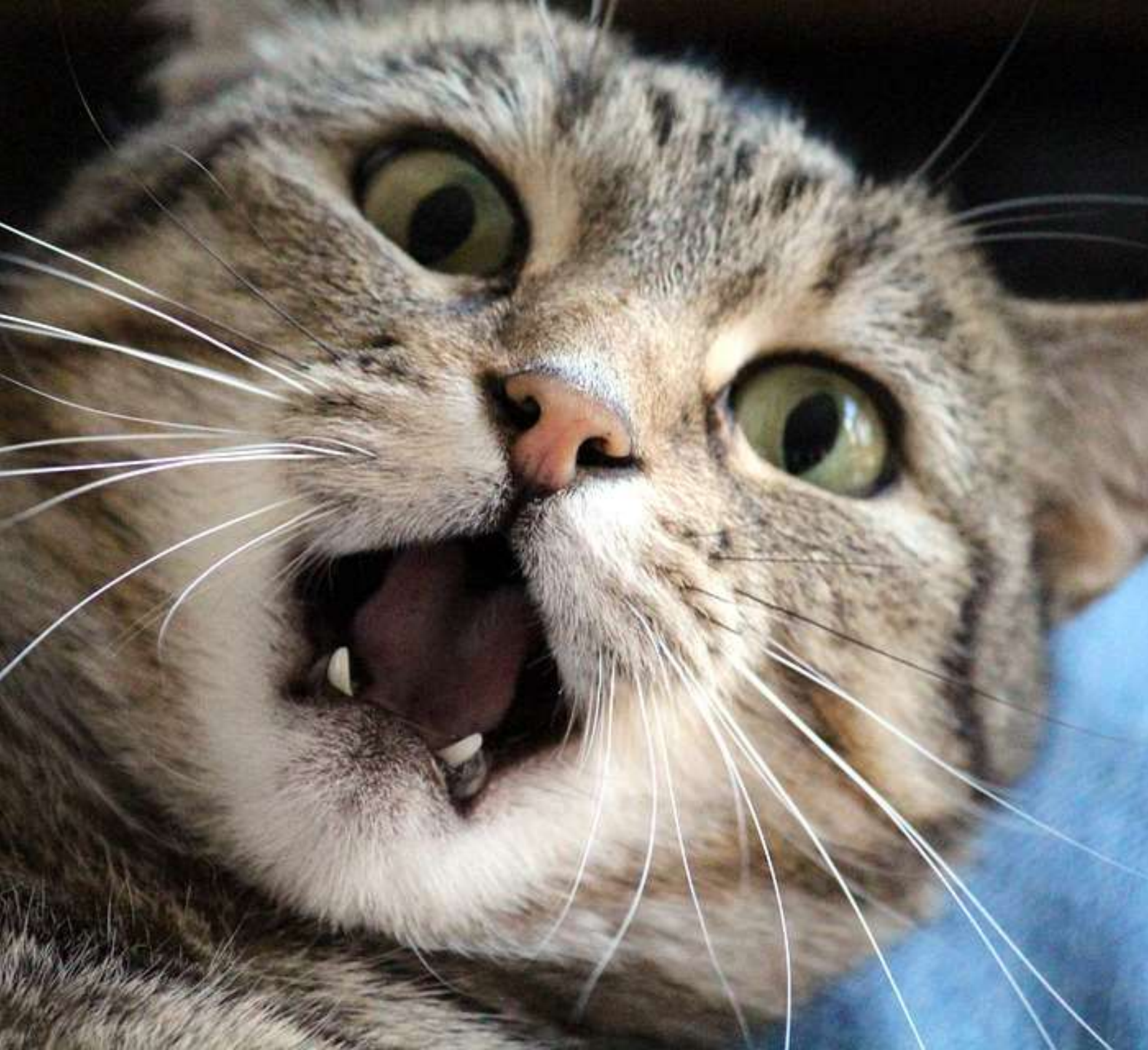
- ProPublica Responds to Company's Critique of Machine Bias Story, ProPublica, Julio 2016.
<https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story>
- Technical Response to Northpointe, ProPublica, Julio 2016.
<https://www.propublica.org/article/technical-response-to-northpointe>
- [Annotated responses to an academic paper](#) that defended Northpointe's approach, ProPublica, Septiembre 2016.
<https://www.documentcloud.org/documents/3248777-Lowenkamp-Fedprobation-sept2016-0.html>

Otros

Unámonos para evitar la discriminación de los algoritmos que nos gobiernan. MIT Technology Review

<https://www.technologyreview.es/s/7950/unamonos-para-evitar-la-discriminacion-de-los-algoritmos-que-nos-gobiernan>

<https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/>



**¿Qué está
pasando
aquí?**

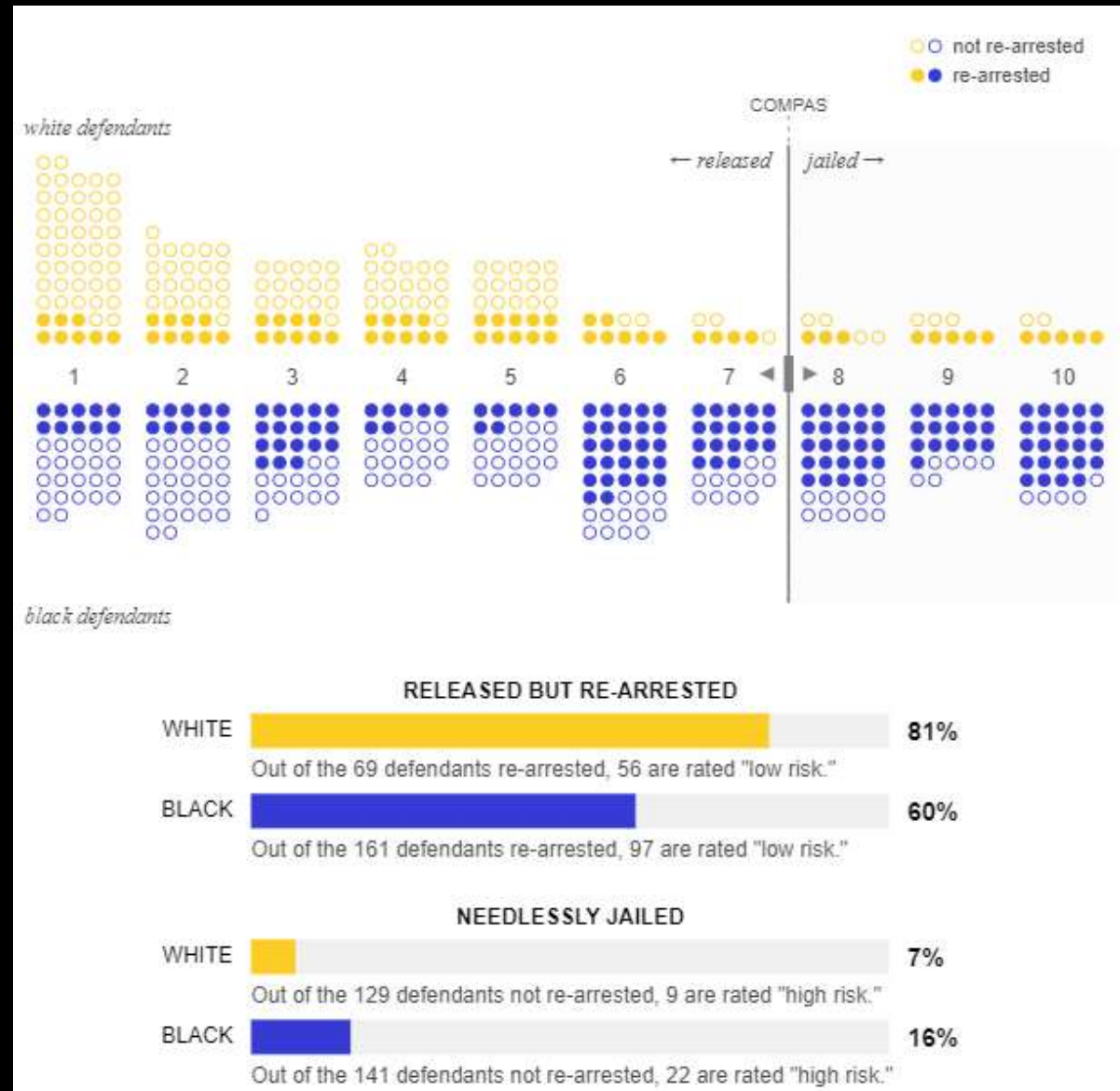
Can you make AI fairer than a judge? Play our courtroom algorithm game

The US criminal legal system uses predictive algorithms to try to make the judicial process less biased. But there's a deeper problem.

by [Karen Hao](#) and [Jonathan Stray](#)

October 17, 2019

<https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/>



Teorema de la imposibilidad de Alexandra Chouldechova

“An instrument that satisfies predictive parity cannot have equal false positive and negative rates across groups when the recidivism prevalence differs across those groups.

[...] Disparate impact can result from the use of a recidivism prediction instrument that is known to satisfy the fairness criterion of predictive parity”.

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

Alexandra Chouldechova. Last revised: February 2017

<https://arxiv.org/pdf/1703.00056.pdf>

¿Qué entendemos entonces por justicia o equidad?

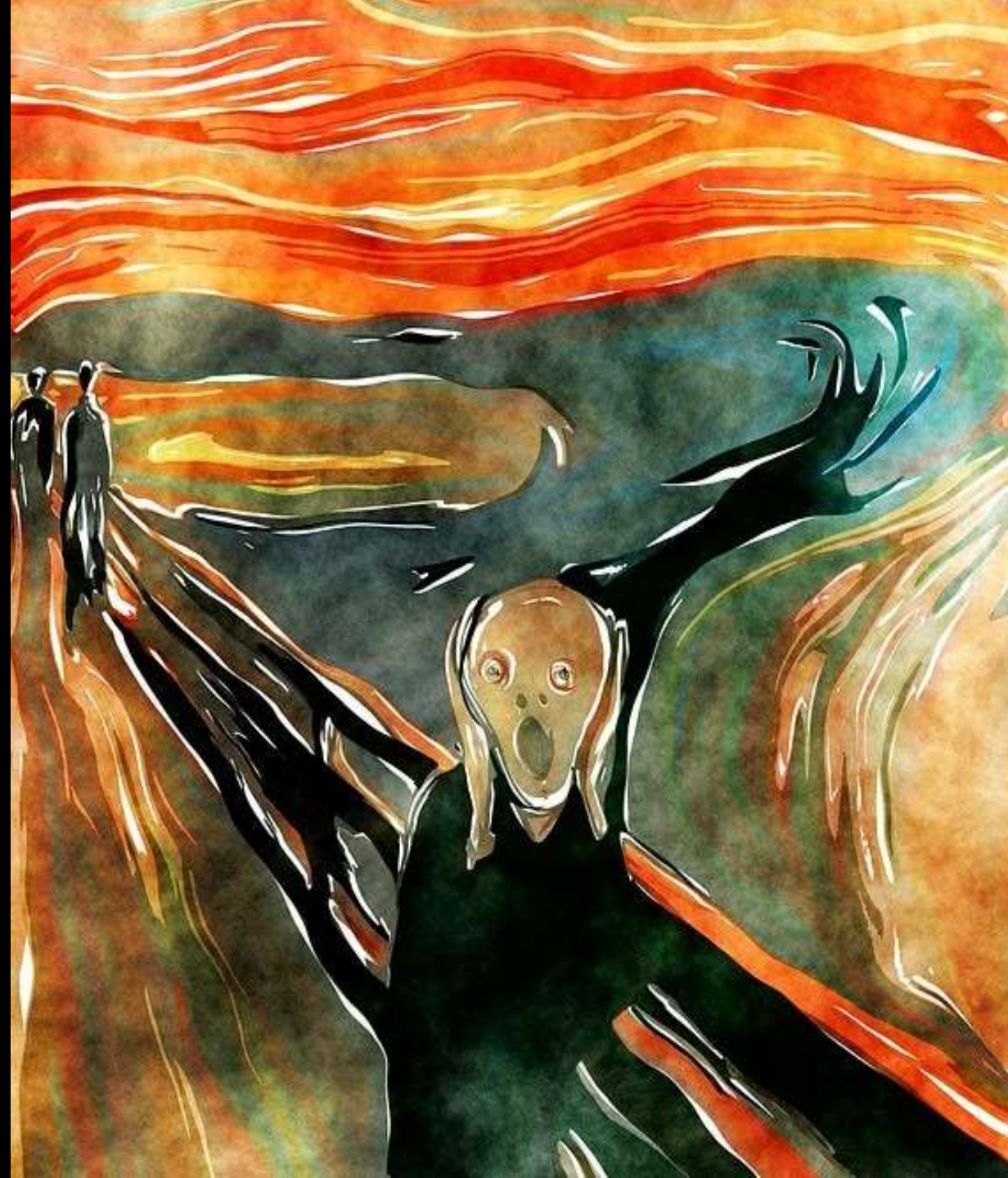
Equidad en los aciertos vs Equidad en los errores.

Diferentes perspectivas:

- La de los jueces: Detectar correctamente la mayor cantidad posible de casos positivos (detenidos reincidentes) es decir, PPV o positive predictive value
- La de los acusados: No ser mantenido en prisión innecesariamente, es decir, FPV o false positive rate.

Análisis

7. ¿Qué tipo de métrica te parece que satisface mejor el principio de justicia en este tipo de casos?
¿Mantendrías el predictive parity o lo sacrificarías para igualar las tasas de error entre grupos?
8. ¿Crees que debe utilizarse la IA en este contexto? ¿La oportunidad de corregir el racismo del sistema penal norteamericano compensa el daño que producen los inevitables errores del modelo?
9. ¿Cómo mejorarías el sistema? ¿Ves otra forma de utilizar un modelo de este tipo que mitigue el impacto sobre los grupos protegidos?



Puntos de vista

"It is difficult to construct a score that doesn't include items that can be correlated with race — such as poverty, joblessness and social marginalization. "If those are omitted from your risk assessment, accuracy goes down,"

"I wanted to stay away from the courts," Brennan said, explaining that his focus was on reducing crime rather than punishment. "But as time went on I started realizing that so many decisions are made, you know, in the courts. So I gradually softened on whether this could be used in the courts or not."

Tim Brennan, PhD, lead developer of the COMPAS y fundador de Northpointe

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Los algoritmos penales "ofrecen la oportunidad de reformar las condenas y revertir el encarcelamiento masivo de una manera científica". Los autores [Anthony Flores, Christopher Lowenkamp, y Kristin Bechtel] temen que esta oportunidad "se esté desvaneciendo debido a la desinformación y el mal entendimiento" que rodea a la tecnología.


<https://www.technologyreview.es/s/7950/unamonos-para-evitar-la-discriminacion-de-los-algoritmos-que-nos-gobiernan>

"The whole point of due process is accuracy, to prevent people from being falsely accused," says Danielle Citron, law professor at the University of Maryland. "The idea that we are going to live with a 40% inaccurate result, that is skewed against a subordinated group, to me is a mind-boggling way to think about accuracy."

<https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story>

"Risk assessments should be impermissible unless both parties get to see all the data that go into them," said Christopher Slobogin, director of the criminal justice program at Vanderbilt Law School. "It should be an open, full-court adversarial proceeding."

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

The background of the slide is a grid of numerous small, rounded-square images of people's faces. The faces are diverse in age, gender, and ethnicity, representing a wide range of human diversity. The grid is arranged in approximately 5 rows and 10 columns, with some faces partially obscured by the text overlay.

“Data biases are inevitable in real world, you must design algorithm to account for them [...]. We need to reframe the problem and move away from mathematical correctness and thus the real challenge is how we make algorithmic systems support human values”.

21 fairness definition and their politics by Arvind Narayanan
(Associate professor at computer science at Princeton University)