The Tyranny of Algorithmic Bias

& How to End It

HELLO!

MATTHEW FINNEY

Data Scientist, Harvard

mattfinney.github.io

AFROTECH WORLD

AI in the 2010s

Sign in
Contribute →
The Guardian

The Washington Post
Democracy Dies in Darkness

HUD is reviewing Twitter's and Google's ad practices as part of housing discrimination pro

Facebook charged with housing discrimination by HUD

The Dept. of Housing and Urban Development charged Facebook March 28 with violating the Fair Housing Act. (Reuters)

By Tracy Jan and Elizabeth Dwoskin

March 28, 2019 at 6:59 p.m. EDT

The Trump administration delivered its first sanction of a tech giant Thursday, charging Facebook with housing discrimination in a move that could threaten the way the industry makes its profits.

BLOOMBERG

ALGORITHM WATCH

story

Google apologizes after its Visi
produced racist results

Published: April 7, 2020

Category: story

By Nicolas Kayser-Bril • nkb@algorithmwatc

A Google service that automatically labels i
starkly different results depending on skin
image. The company fixed the issue, but th
much broader.

There is a saying in computer science: garbag
garbage out. When we feed machines data th
our prejudices, they mimic them – from anti
chatbots to racially biased software. Does a h
future await people forced to live at the mercy of
algorithms?

Antuan Goodwin
Sept. 23, 2020 7:53 a.m. PT

34

The New York Times

A.I. Is Learning to Read Mammograms

Computers that are trained to recognize patterns and interpret
humans at finding cancer on X-rays.

The New York Times

Surveillance cameras have been deployed across Detroit as part of Project Green Light, which is meant to deter crime. Brittany Greeson for The New York Times

As Cameras Track Detroit's Residents, a Debate Ensues Over Racial Bias

Studies have shown that facial recognition software can return more false matches for African-Americans than for white people, a sign of what experts call "algorithmic bias."

By Amy Harmon

July 8, 2019

The Socially Conscious Data Scientist's Agenda

1. We can **define and measure** algorithmic bias

2. We can **isolate the root cause** of (poor) algorithmic behavior

3. We can **take action** to make algorithms more fair

What is algorithmic bias?

Case study

In the U.S., kidney function measurements are adjusted by race

- The eGFR is the standard-of-care for measuring kidney function
- It's calculated by measuring the level of creatinine in a blood sample
- Because "African Americans" have higher muscle mass, the CKD-EPI algorithm increases their scores
- A higher score indicates higher kidney function

The CKD-EPI eGFR equation is racially biased

HAS
HAUTE AUTORITÉ DE SANTÉ

TEXTE COURT DU RAPPORT D'EVALUATION TECHNOLOGIQUE

Evaluation du débit de filtration glomérulaire
et du dosage de la créatininémie
dans le diagnostic de la maladie rénale chronique
chez l'adulte

Décembre 2011

Service évaluation des actes profi

EXTRA nephron

Nephron Extra 2012;2:293–302
DOI: 10.1159/000343899
Published online: December 1, 2012

Original Paper

Race Adjustment for Estimating Glomerular Filtration Rate Is Not Always Necessary

Juliana A. Zanocco Sonia K. Nishida Michelle Tiveron Passos
Amélia Rodrigues Pereira Marcelo S. Silva
Aparecido B. Pereira Gianna Mastroianni Kirsztajn

Glomerulopathy Section, Division of Nephrology, Medicine Department, Federal University of São Paulo (UNIFESP), São Paulo, Brazil

ltration rate · Glomerulonephritis

te (eGFR) is very important in clinical practice, populations. We aimed at establishing the best nphasis on the need for race correction. ***Meth-*** idney disease (CKD) and 42 without previously ned by urinalysis. Serum creatinine and plasma GFR was estimated by the Mayo Clinic, abbrevi- RD) and Chronic Kidney Disease Epidemiology e clearance was estimated by the Cockcroft- was used as the gold standard for GFR determi- formula (BreGFR). ***Results:*** Measured and esti- 7% female, with a mean age of 41 years (range clients among the plasma clearance of iohexol) and corresponded to the following scores: CG 304; MDRD 0.848; MDRD1 (without race adjust- ace adjustment) 0.876, and BreGFR 0.844. ***Con-***

ction, Federal University of São Paulo (UNIFESP)

-900 (Brazil)
o.unifesp.br

KARGER

ELSEVIER

European Journal of Obstetrics & Gynecology and Reproductive Biology
Volume 176, May 2014, Pages 200-201

LETTER TO THE EDITOR

Adjustment for race in the estimation of glomerular filtration rate (GFR) is inappropriate in the British postnatal population

Anna L. Roberts, Alastair Ferraro, Amanda Green, Pam Loughna, Fiona Broughton-Pipkin

The CKD-EPI eGFR equation is racially biased

HAS
HAUTE AUTORITÉ DE SANTÉ

TEXTE COURT DU RAPPORT D'EVALUATION TECHNOLOGIQUE

Evaluation du débit de filtration glomérulaire
et du dosage de la créatininémie
dans le diagnostic de la maladie rénale chronique
chez l'adulte

Décembre 2011

Service évaluation des actes profi

EXTRA nephron

Nephron Extra 2012;2:293–302
DOI: 10.1159/000343899
Published online: December 1, 2012

© 2012 S. Karger AG, Basel
www.karger.com/nne

293

Original Paper

Race Adjustment for Estimating Glomerular Filtration Rate Is Not Always Necessary

Juliana A. Zanocco Sonia K. Nishida Michelle Tiveron Passos
Amélia Rodrigues Pereira Marcelo S. Silva
Aparecido B. Pereira Gianna Mastroianni Kirsztajn

Glomerulopathy Section, Division of Nephrology, Medicine Department, Federal University of São Paulo (UNIFESP), São Paulo, Brazil

KARGER

ELSEVIER

European Journal of Obstetrics & Gynecology and Reproductive Biology
Volume 176, May 2014, Pages 200-201

LETTER TO THE EDITOR

Adjustment for race in the estimation of glomerular filtration rate (GFR) is inappropriate in the British postnatal population

Anna L. Roberts, Alastair Ferraro, Amanda Green, Pam Loughna, Fiona Broughton-Pipkin

Many people see this as unfair. Can you think of any reasons why?

What is fairness?

Two definitions used in the algorithmic community

Group Fairness

Identifiable groups should be treated similarly to the population as a whole

Individual Fairness

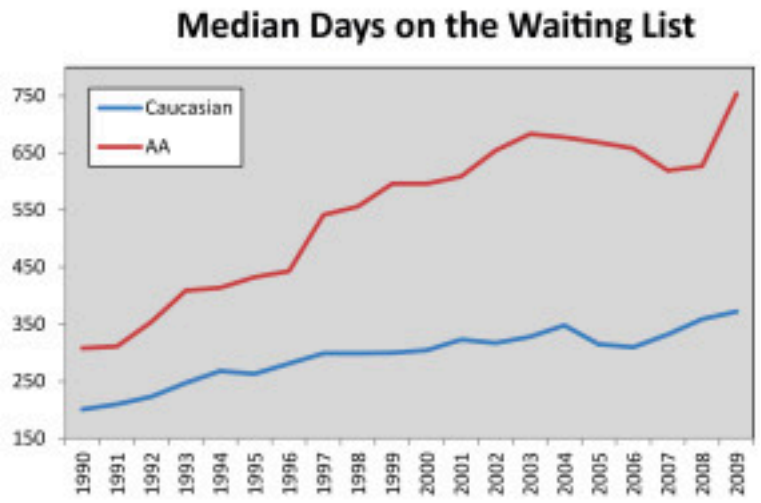
Similar individuals should be treated similarly

Adapted from Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. https://fairware.cs.umass.edu/papers/Verma.pdf.

Is the CKD-EPI algorithm Group Fair?

Group Fairness Definition

Protected groups should be treated similarly to non-protected groups and the population as a whole

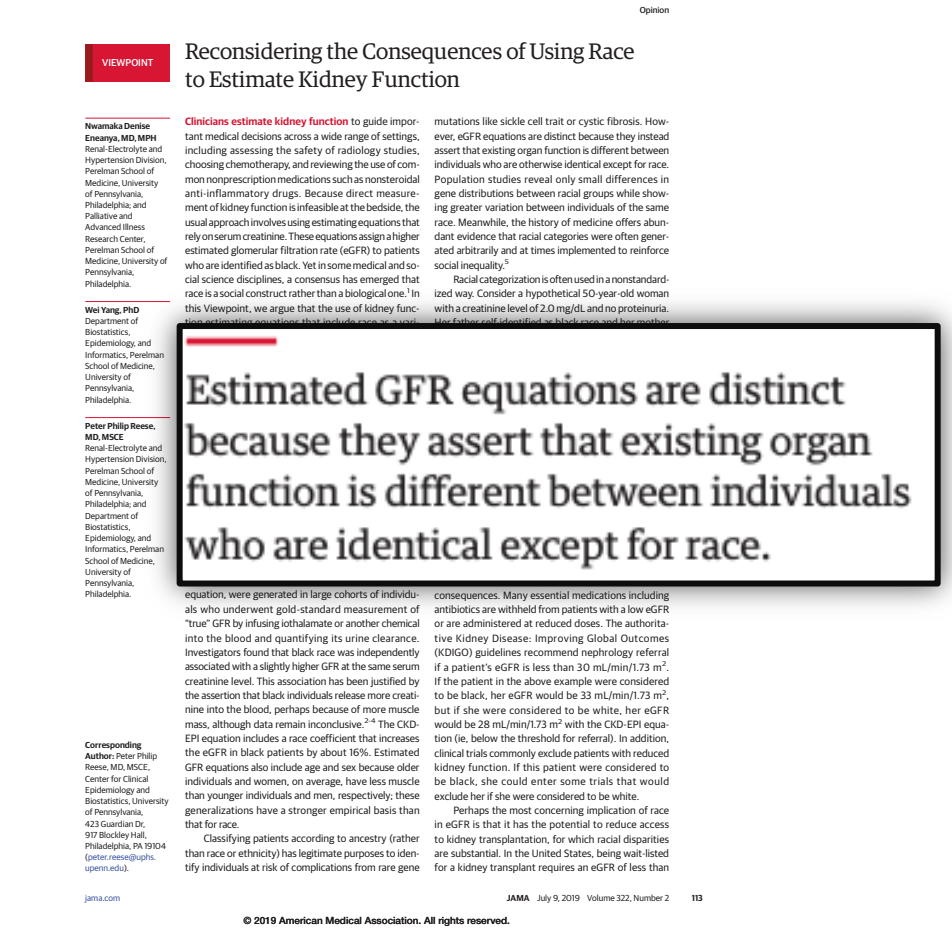


Source: Taber et al., Twenty years of evolving trends in racial disparities for adult kidney transplant recipients. Kidney Int. 2016.

Is the CKD-EPI algorithm Individually Fair?

Individual Fairness Definition

Similar individuals should be treated similarly



Opinion

VIEWPOINT

Reconsidering the Consequences of Using Race to Estimate Kidney Function

Nwamaka Denise Eneanya, MD, MPH
Renal-Electrolyte and Hypertension Division, Perelman School of Medicine, University of Pennsylvania, Philadelphia; and Palliative and Advanced Illness Research Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia.

Wei Yang, PhD
Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia.

Peter Philip Reese, MD, MSCE
Renal-Electrolyte and Hypertension Division, Perelman School of Medicine, University of Pennsylvania, Philadelphia; and Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia.

Corresponding Author: Peter Philip Reese, MD, MSCE, Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, 423 Guardian Dr, 917 Blockley Hall, Philadelphia, PA 19104 (peter.reese@uphs.upenn.edu).

Clinicians estimate kidney function to guide important medical decisions across a wide range of settings, including assessing the safety of radiology studies, choosing chemotherapy, and reviewing the use of common nonprescription medications such as nonsteroidal anti-inflammatory drugs. Because direct measurement of kidney function is infeasible at the bedside, the usual approach involves using estimating equations that rely on serum creatinine. These equations assign a higher estimated glomerular filtration rate (eGFR) to patients who are identified as black. Yet in some medical and social science disciplines, a consensus has emerged that race is a social construct rather than a biological one.[1] In this Viewpoint, we argue that the use of kidney function estimating equations that include race as a vari-

equation, were generated in large cohorts of individuals who underwent gold-standard measurement of "true" GFR by infusing iothalamate or another chemical into the blood and quantifying its urine clearance. Investigators found that black race was independently associated with a slightly higher GFR at the same serum creatinine level. This association has been justified by the assertion that black individuals release more creatinine into the blood, perhaps because of more muscle mass, although data remain inconclusive.[2-4] The CKD-EPI equation includes a race coefficient that increases the eGFR in black patients by about 16%. Estimated GFR equations also include age and sex because older individuals and women, on average, have less muscle than younger individuals and men, respectively; these generalizations have a stronger empirical basis than that for race.

Classifying patients according to ancestry (rather than race or ethnicity) has legitimate purposes to identify individuals at risk of complications from rare gene mutations like sickle cell trait or cystic fibrosis. However, eGFR equations are distinct because they instead assert that existing organ function is different between individuals who are otherwise identical except for race. Population studies reveal only small differences in gene distributions between racial groups while showing greater variation between individuals of the same race. Meanwhile, the history of medicine offers abundant evidence that racial categories were often generated arbitrarily and at times implemented to reinforce social inequality.[5]

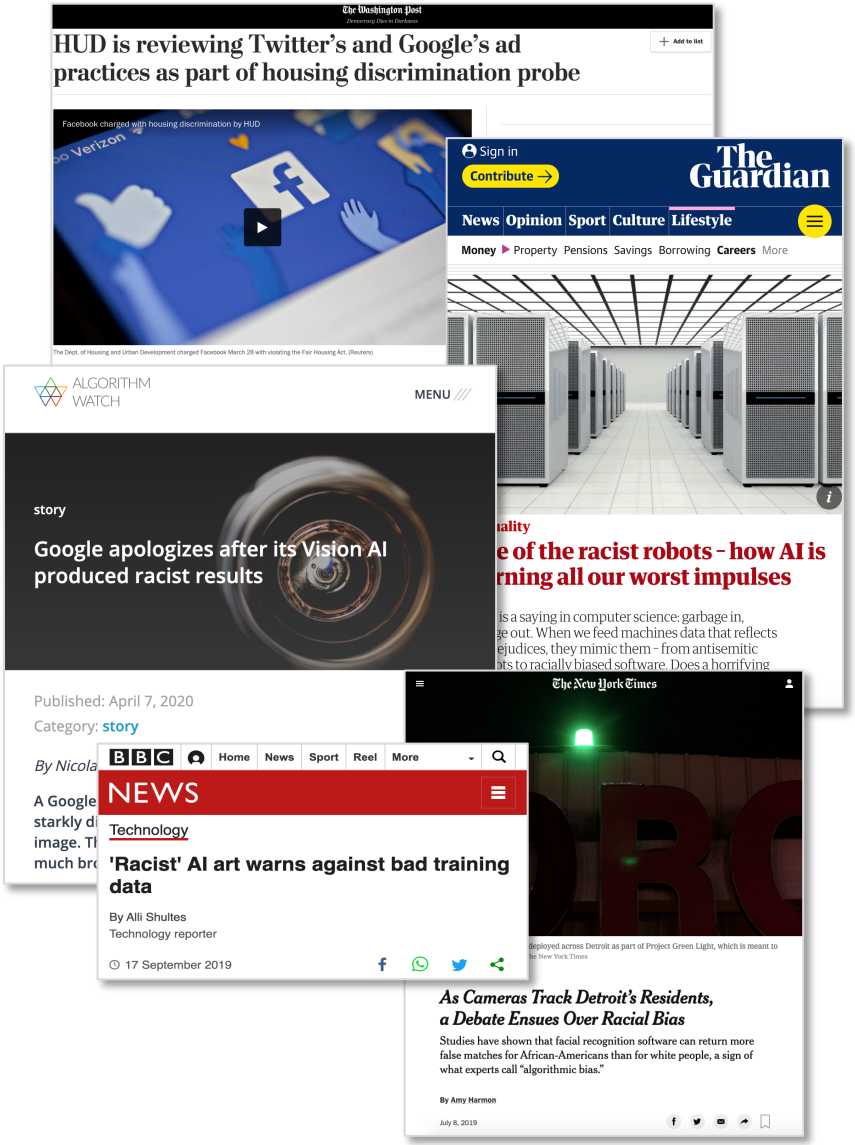
Racial categorization is often used in a nonstandardized way. Consider a hypothetical 50-year-old woman with a creatinine level of 2.0 mg/dL and no proteinuria. Her father self-identified as black race and her mother

consequences. Many essential medications including antibiotics are withheld from patients with a low eGFR or are administered at reduced doses. The authoritative Kidney Disease: Improving Global Outcomes (KDIGO) guidelines recommend nephrology referral if a patient's eGFR is less than 30 mL/min/1.73 m^2. If the patient in the above example were considered to be black, her eGFR would be 33 mL/min/1.73 m^2, but if she were considered to be white, her eGFR would be 28 mL/min/1.73 m^2 with the CKD-EPI equation (ie, below the threshold for referral). In addition, clinical trials commonly exclude patients with reduced kidney function. If this patient were considered to be black, she could enter some trials that would exclude her if she were considered to be white.

Perhaps the most concerning implication of race in eGFR is that it has the potential to reduce access to kidney transplantation, for which racial disparities are substantial. In the United States, being wait-listed for a kidney transplant requires an eGFR of less than

Estimated GFR equations are distinct because they assert that existing organ function is different between individuals who are identical except for race.

Why does this keep happening?



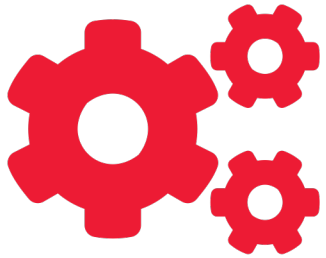
How do we make models?

TECHNOLOGY

- Data
- Algorithm
- …

PEOPLE

- Data Analyst/Scientist
- Business Owner
- End User
- …



PROCESS

- Model Training
- Evaluation
- Application
- …

How did we make a biased kidney function model?

TECHNOLOGY

The CKD-EPI regression was selected among other viable measures

How did we make a biased kidney function model?

TECHNOLOGY

The CKD-EPI regression was selected among other viable measures

PEOPLE

We'll assume the researchers' best intentions

How did we make a biased kidney function model?

TECHNOLOGY

The CKD-EPI regression was selected among other viable measures

PEOPLE

We'll assume the researchers' best intentions

PROCESS

The process was optimized for overall accuracy

How did we make a biased kidney function model?

TECHNOLOGY

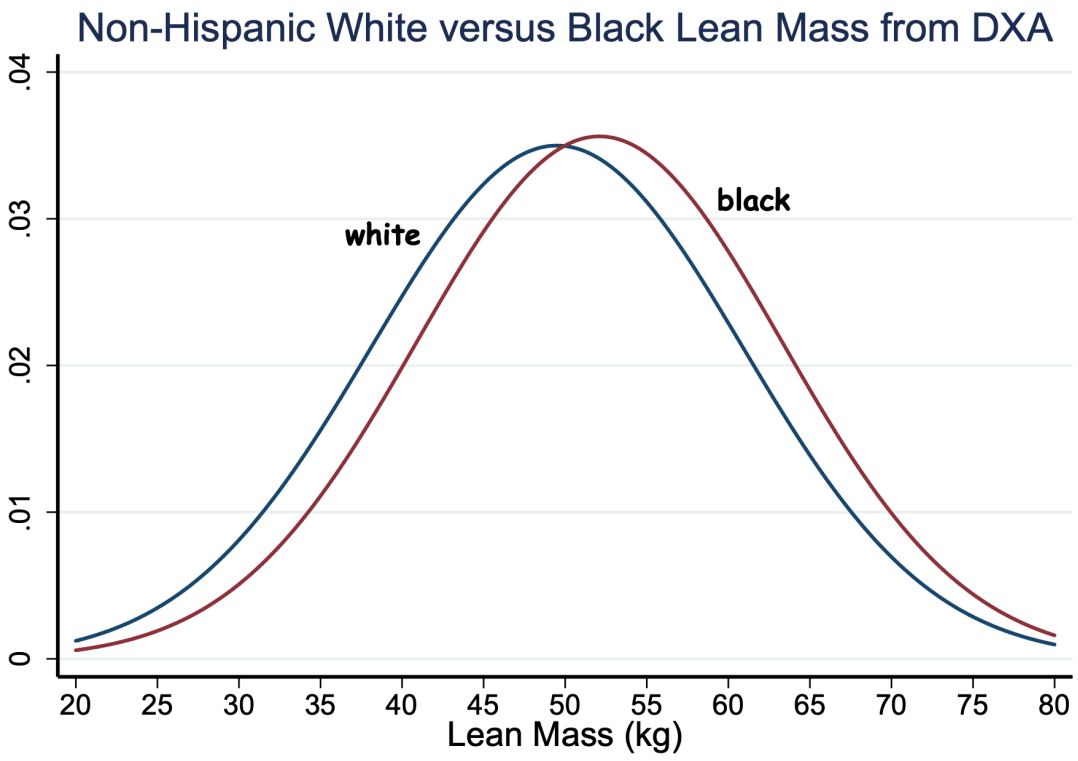
The CKD-EPI regression was selected among other viable measures

PEOPLE

We'll assume the researchers' best intentions

PROCESS

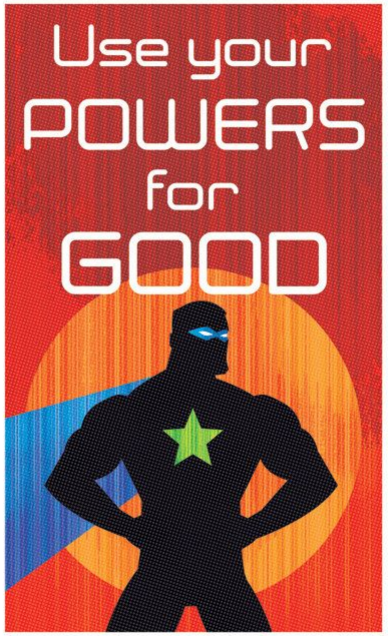
The process was optimized for overall accuracy



Source: National Health and Nutrition Examination Survey

Why isn't fairness part of our process?

We have good intentions



… but need mechanisms for action

CHALLENGES

- Hard to define
- Hard to measure

Fairness is context-specific

(LACK OF) INCENTIVES

- Lack of transparency
- Lack of accountability

No hard business reason to prioritize fairness

Image: Rob Osborne

How will we end this?

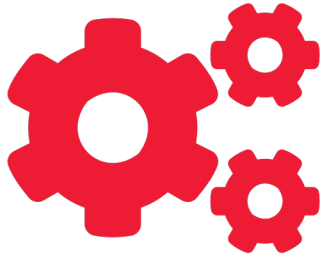
Ingredients of an algorithmic decision

TECHNOLOGY

- Data
- Algorithm
- …

PEOPLE

- Data Analyst/Scientist
- Business Owner
- End User
- …

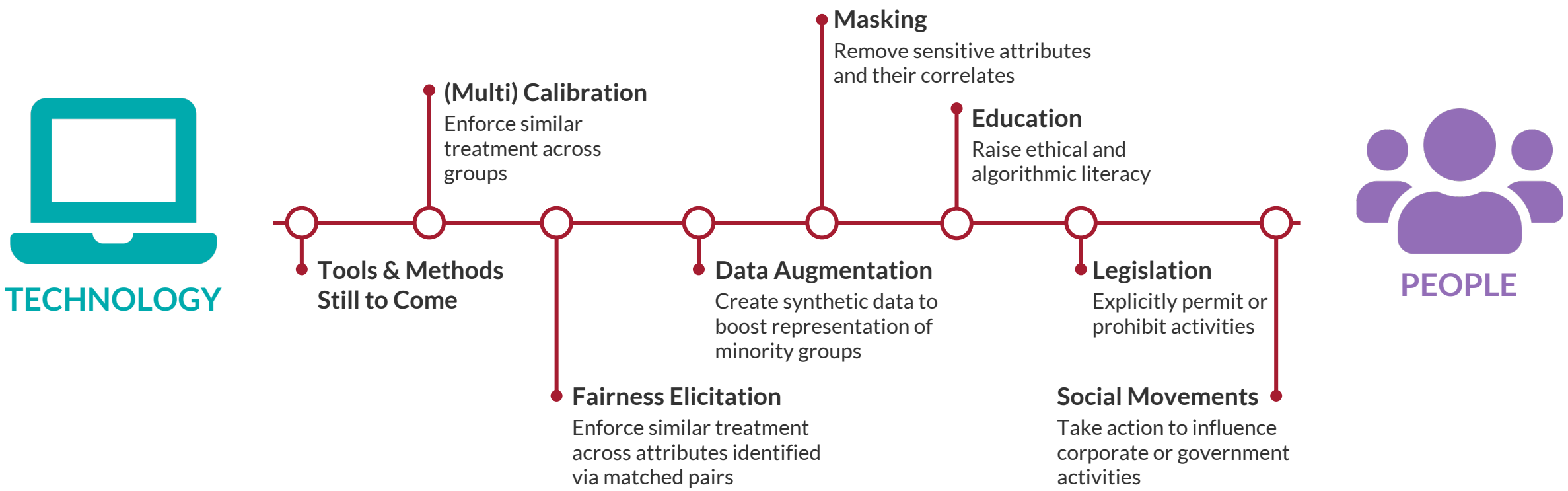


PROCESS

- Model Training
- Evaluation
- Application
- …

How can we change these to mitigate algorithmic bias?

Existing approaches focus on the Technology and People axes



Existing approaches focus on the Technology and People axes

TECHNOLOGY

IBM Research Trusted AI

AI Fairness 360

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it.

Python API Docs ↗ | Get Python Code ↗ | Get R Code ↗

Not sure what to do first? Start here!

Read More

Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.

→

Try a Web Demo

Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.

→

Watch Videos

Watch videos to learn more about AI Fairness 360.

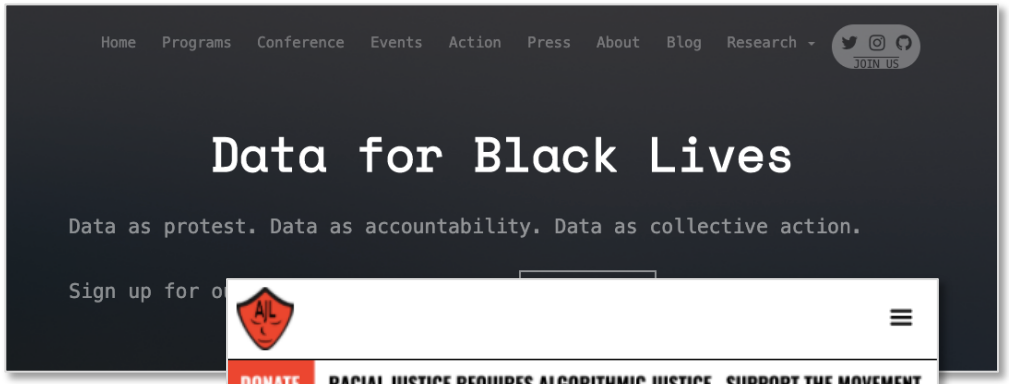
→

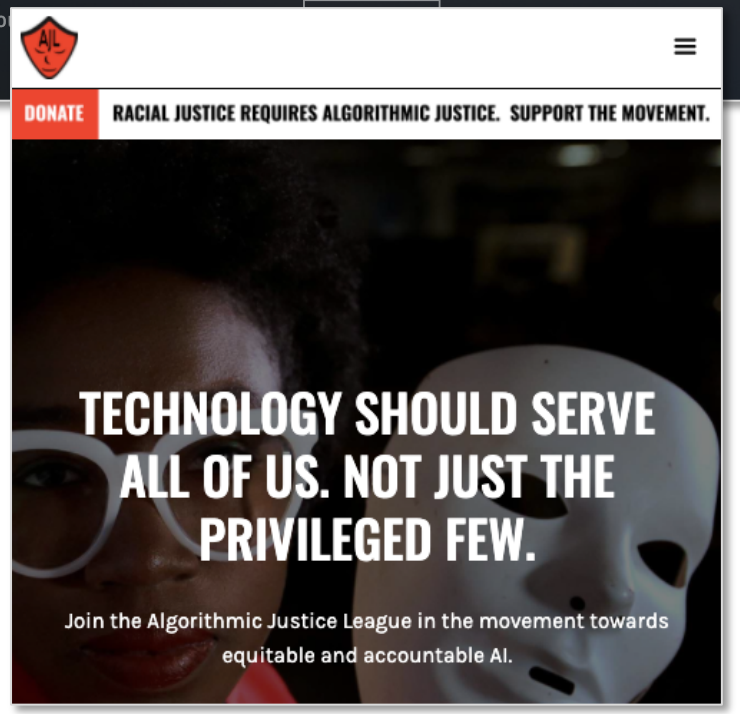
Symposium on Foundations of Responsible Computing (FORC)

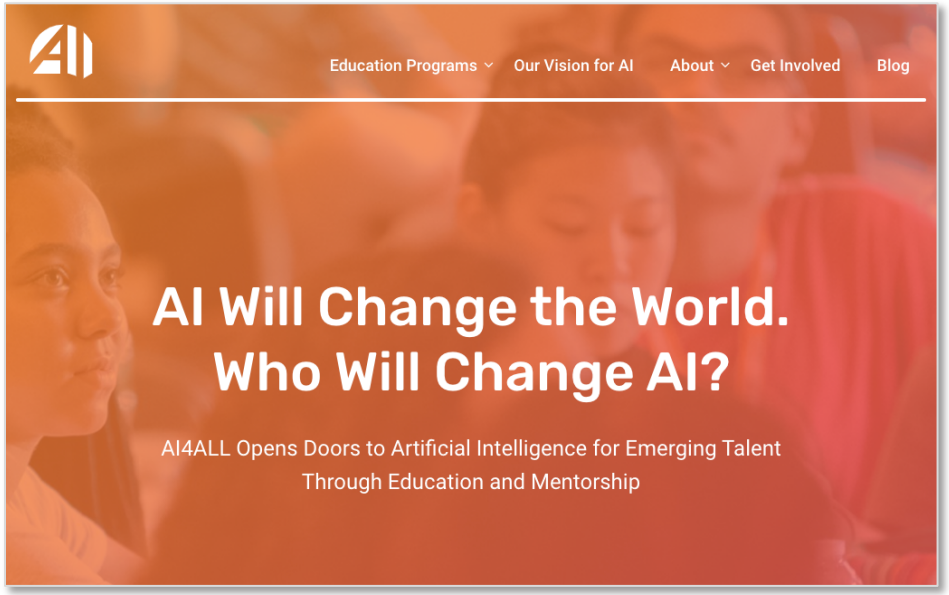
Current Academic Year Events, Events - Comments Off

The Symposium on Foundations of Responsible Computing (FORC) is a forum for mathematical research in computation and society writ large. The Symposium aims to catalyze the formation of a community supportive of the application of theoretical computer science, statistics, economics and other relevant analytical fields to problems of pressing and anticipated societal concern.

Existing approaches focus on the Technology and People axes







PEOPLE

Fixing our Process to realize algorithmic fairness intentions

What mechanisms can help us build fair models?

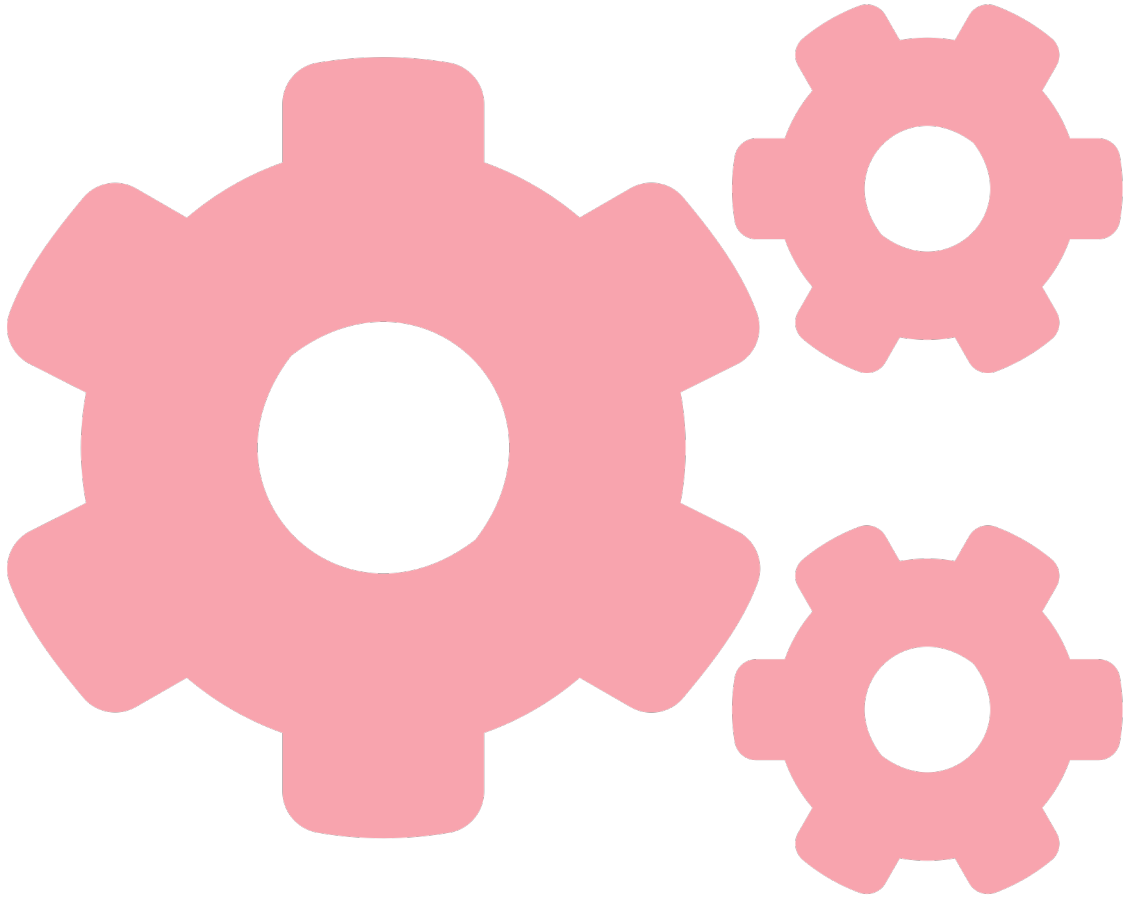
CHALLENGES

Hard to define

Hard to measure

Lack of transparency

Lack of accountability



Fixing our Process to realize algorithmic fairness intentions

What mechanisms can help us build fair models?

CHALLENGES

Hard to define

Hard to measure

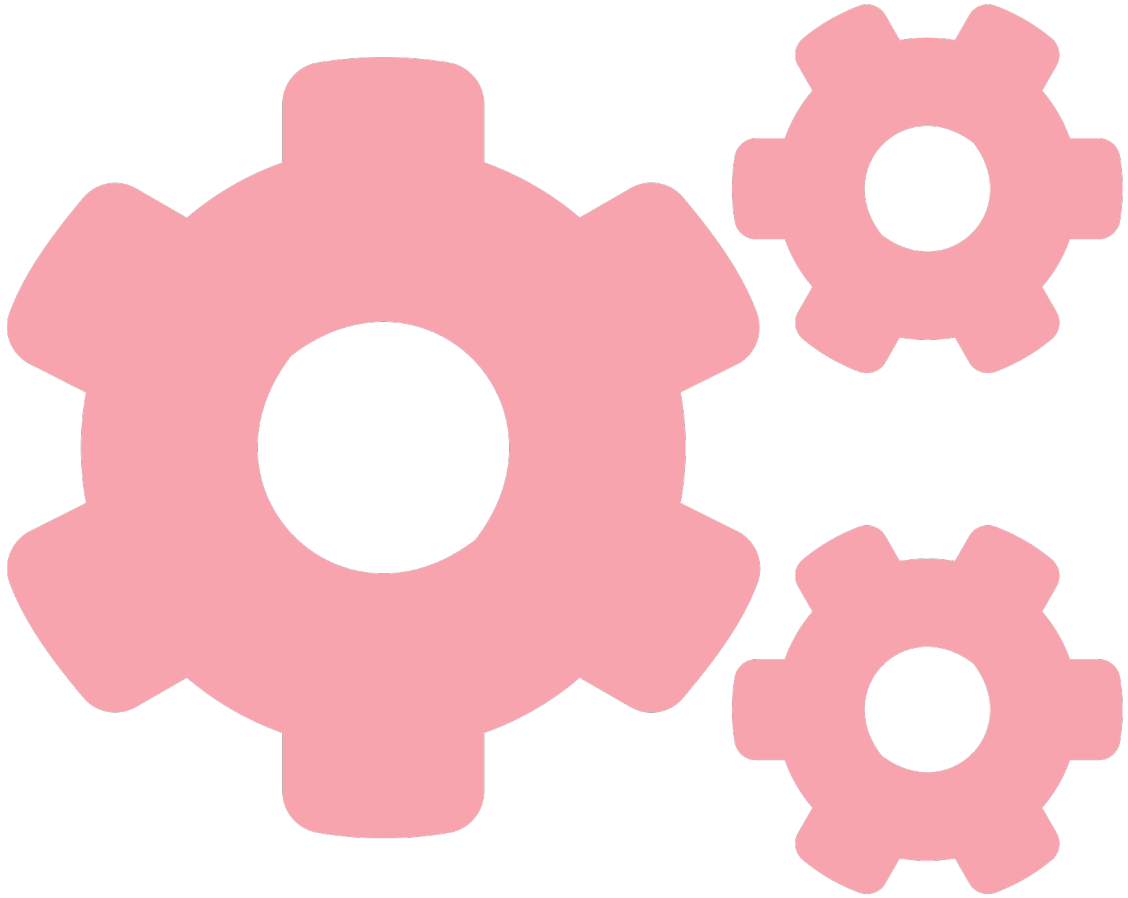
Lack of transparency

Lack of accountability

PROPOSED APPROACH

Fairness Statement

A commitment to defined and measurable fairness objectives



What is a Fairness Statement?

An application-specific commitment to defined and measurable fairness goals

SCOPE

- Define the relevant fairness objective (or constraint) for your application
- Document potential sources of bias as well as the downstream impact to individuals or groups
- Identify appropriate controls (procedural and algorithmic) to mitigate unacceptable risks

BENEFITS

- Work towards a named goal
- Inform choices and tradeoffs in algorithmic development and deployment
- Catch problems early
- Measure your progress/compliance

Fixing our Process to realize algorithmic fairness intentions

What mechanisms can help us build fair models?

CHALLENGES

Hard to define

Hard to measure

Lack of transparency

Lack of accountability

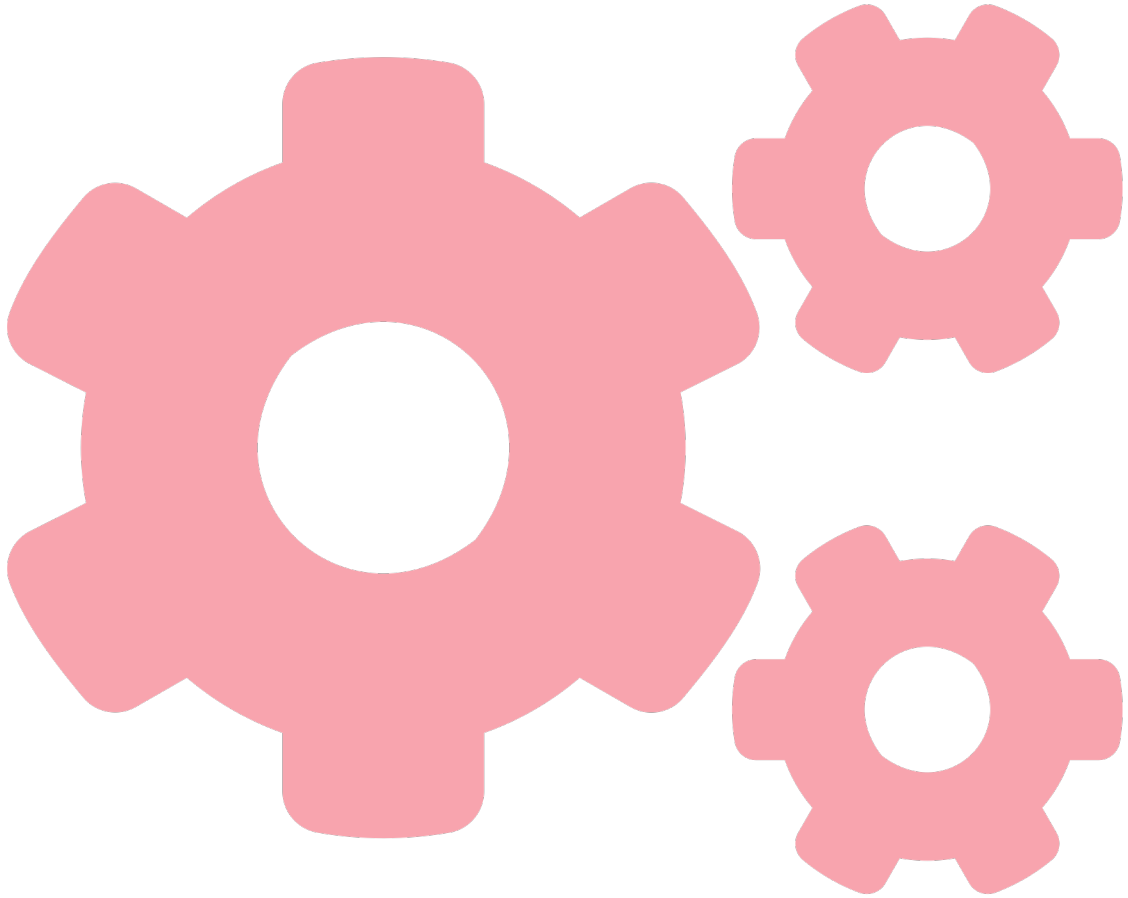
PROPOSED APPROACH

Fairness Statement

A commitment to defined and measurable fairness objectives

Algorithmic Practice Audit

An independent, third party review of processes and outcomes



What is an Algorithmic Practice Audit?

An independent, third party review of an organization's algorithmic processes and outcomes

SCOPE

- Process
  - Is training data representative?
  - Does data cleaning / presentation introduce bias?
  - Are fair classes of algorithms used?
- Outcomes
  - Does the model meet its stated fairness goals?
  - Is there disparate impact or measurable bias?
  - Is bias introduced by humans in the "last mile"?

BENEFITS

- Signal to consumers and (shareholders) that algorithmic services are correct and fair
- Use a forcing function to improve internal processes and controls
- Take pride in certification that you're doing the right thing

Mechanisms work

- In 2014, the Dutch government developed the System Risk Indicator (SyRI) to detect benefit fraud
- Low-income and immigrant neighborhoods were more likely to be targeted
- A court in The Hague shut it down due to discrimination based on their socio-economic status, ethnicity, and religion

Sign in

Support us →

The Guardian

News Opinion Sport Culture Lifestyle

World ▸ Europe US Americas Asia Australia Middle East Africa Inequality

Artificial intelligence (AI)

This article is more than 7 months old

Welfare surveillance system violates human rights, Dutch court rules

Government told to halt use of AI to detect fraud in decision hailed by privacy campaigners

Jon Henley *and* Robert Booth

Wed 5 Feb 2020 08.18 EST

What will you do to create fair algorithms?

	Organizations	Individuals
TECHNOLOGY	Are you following existing technical best practices, and using classes of fair algorithms?	Are you aware of all the algorithmic decisions in your life?
PEOPLE	Are your data and tech teams representative of your customers and stakeholders?	Do you invest in your data literacy skills?
PROCESS	Do you have mechanisms to ensure algorithmic fairness?	Do you request and review algorithmic audits?

Takeaways

1. We can **define and measure** algorithmic bias

2. We can **isolate the root cause** of (poor) algorithmic behavior

3. We can **take action** to make algorithms more fair

Matthew Finney

Data Scientist, Harvard University

mattfinney.github.io

Thank you!