The Flaws of Policies Requiring Human Oversight of Government Algorithms

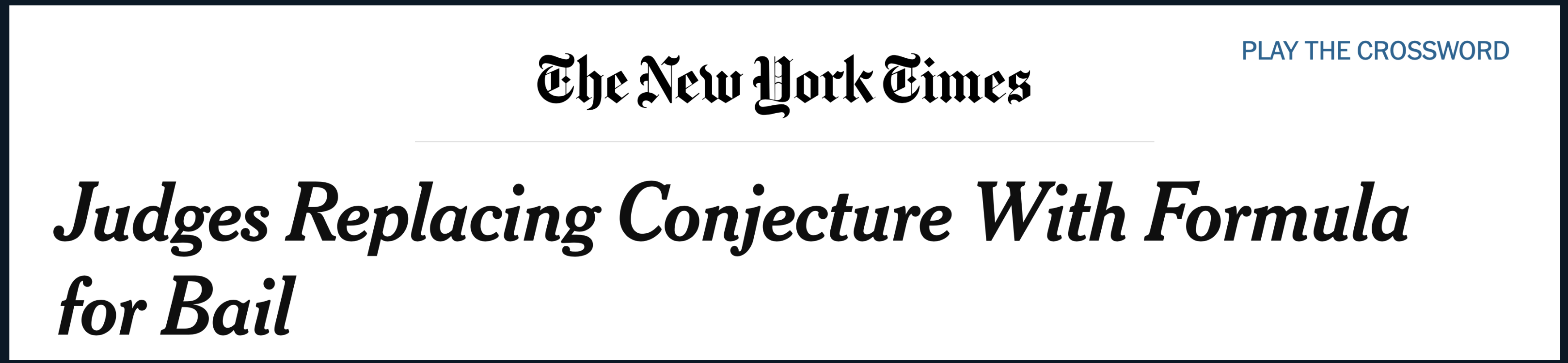
Ben Green

University of Michigan

@benzevgreen
benzevgreen.com
bzgreen@umich.edu

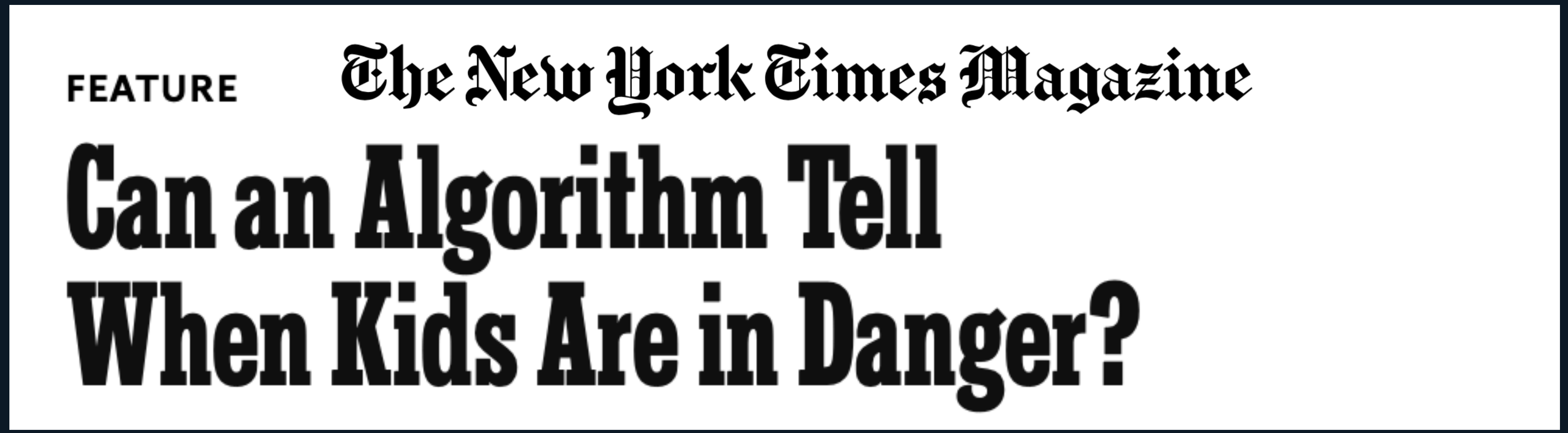
The Promises and Perils of Government Algorithms

Promises: Accuracy, Fairness, Consistency

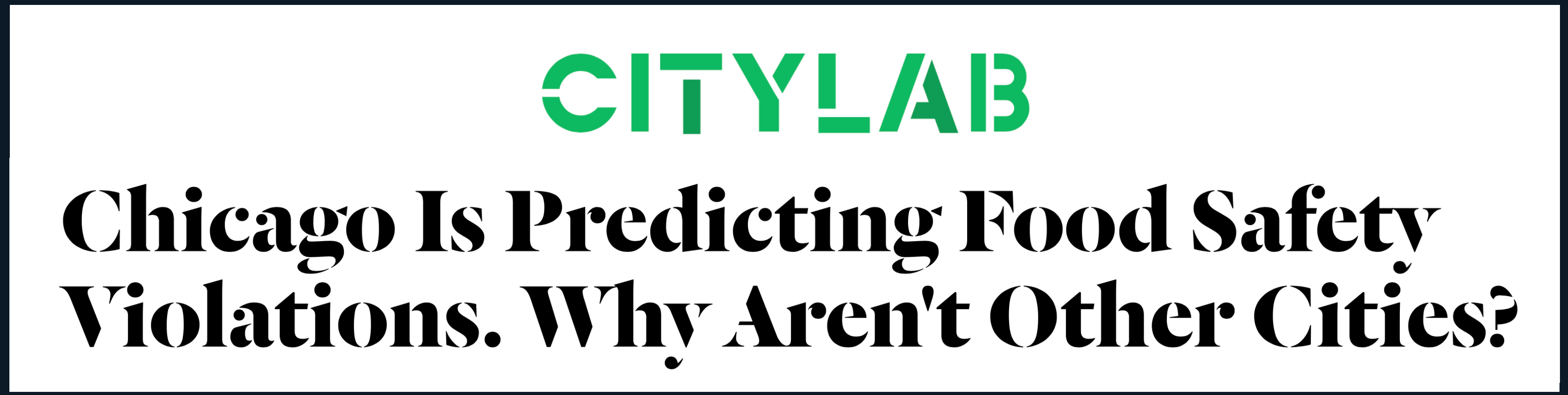

The New York Times

PLAY THE CROSSWORD

Judges Replacing Conjecture With Formula for Bail

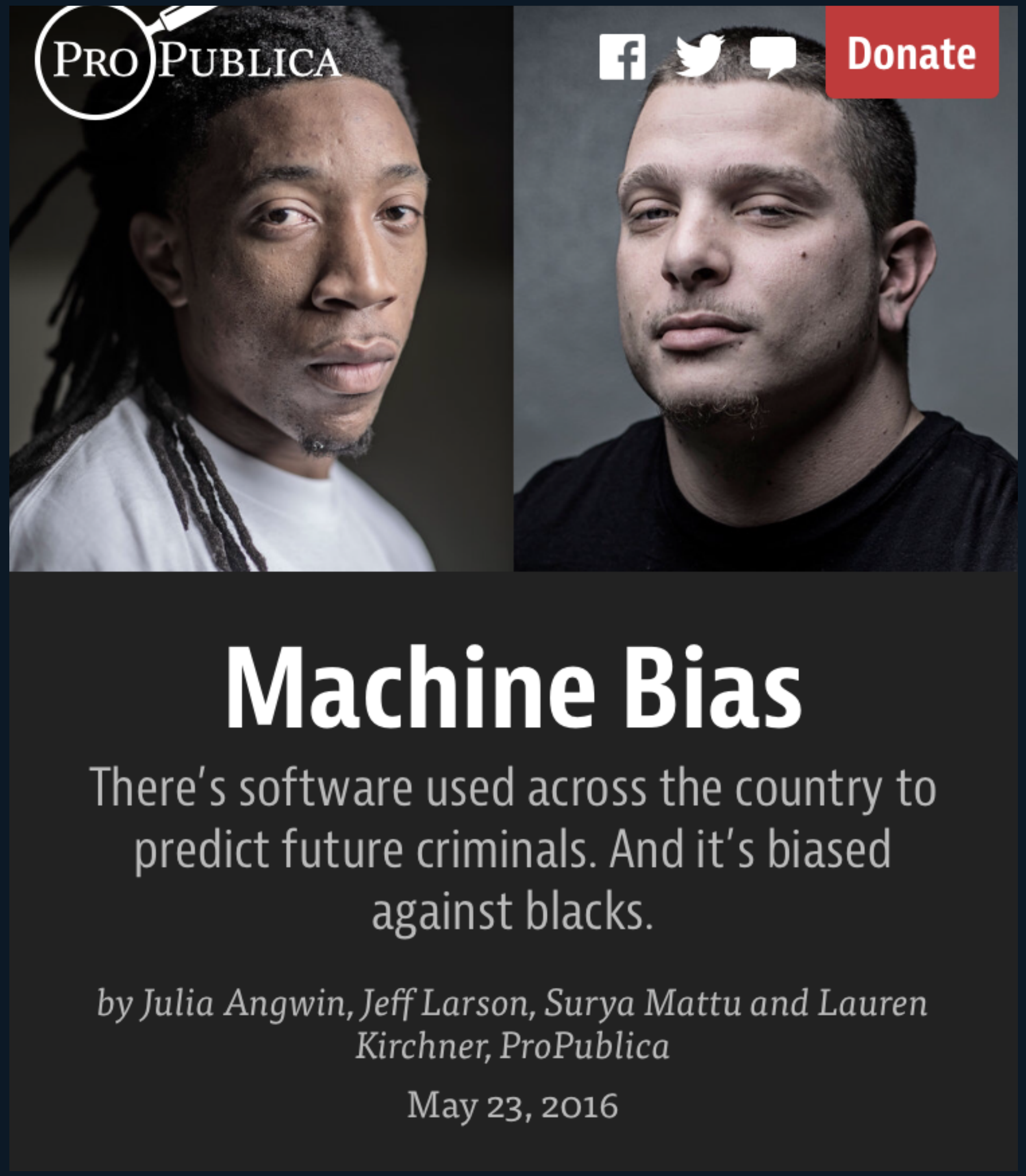

FEATURE The New York Times Magazine

Can an Algorithm Tell When Kids Are in Danger?


CITYLAB

Chicago Is Predicting Food Safety Violations. Why Aren't Other Cities?

Perils: Errors, Biases, and Inflexibility

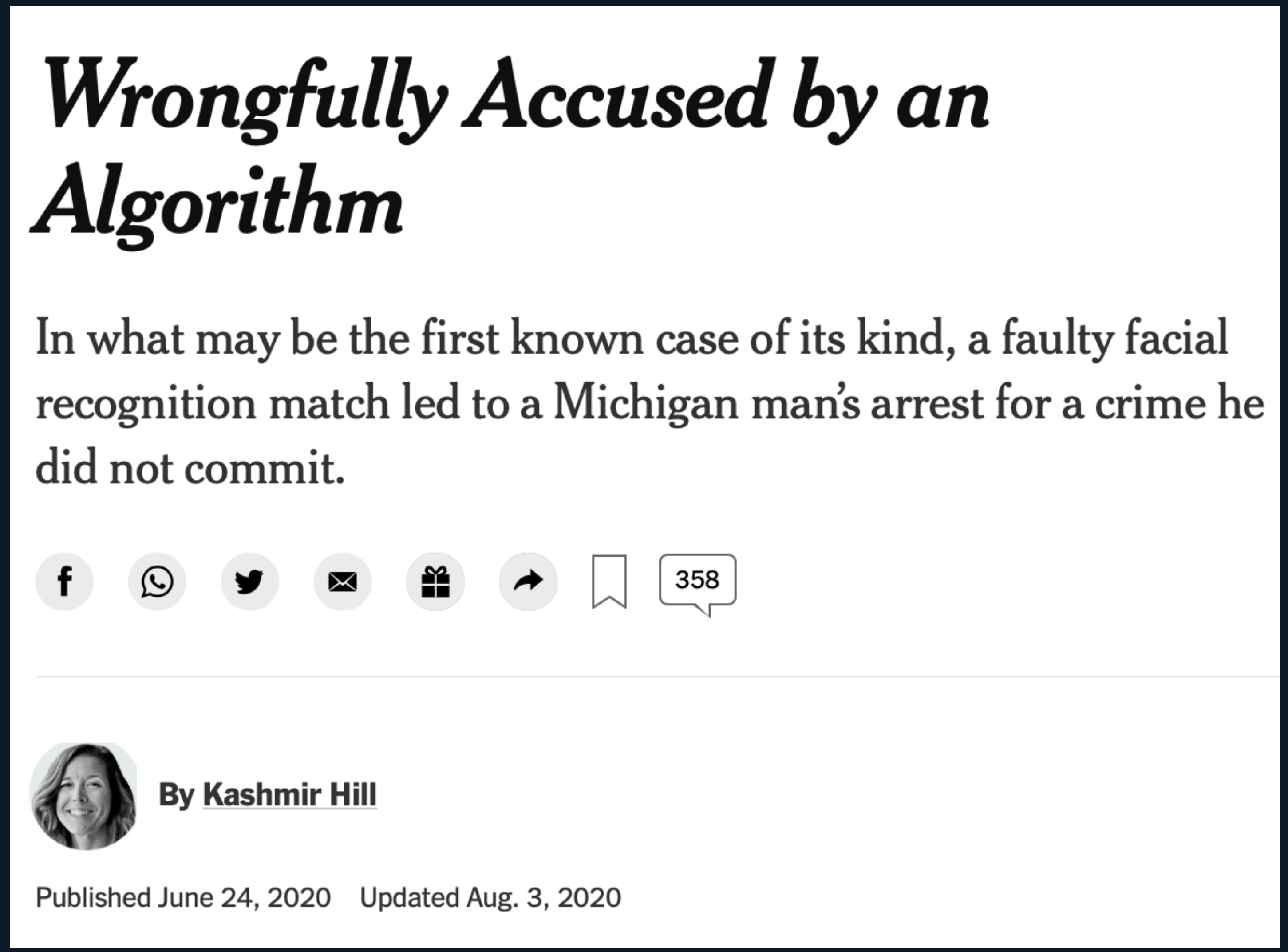


Machine Bias

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


Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

By Kashmir Hill

Published June 24, 2020 Updated Aug. 3, 2020

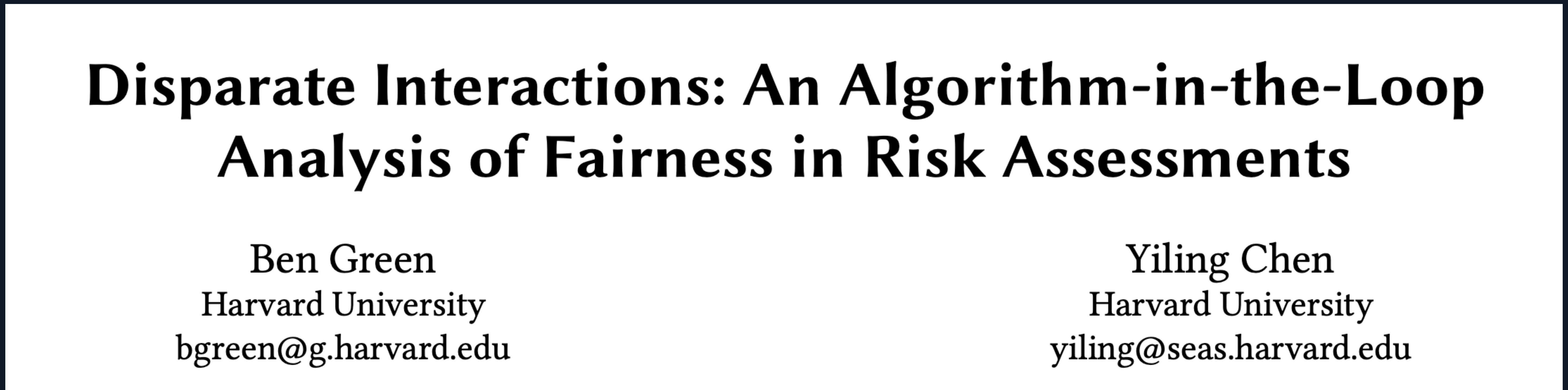
Policy question: How can governments improve public policy using algorithms while preventing the harms of algorithms?

Human Oversight as Algorithmic Regulation

- European Commission AI Act: “For high-risk AI systems, […] human oversight [… is] strictly necessary to mitigate the risks to fundamental rights and safety posed by AI.”
- Canadian Directive on Automated Decision-Making: Decisions likely to have “high” or “very high” social impacts “cannot be made without having specific human intervention points during the decision-making process; and the final decision must be made by a human.”

Prior Work on Human-Algorithm Collaboration

FAT* 2019

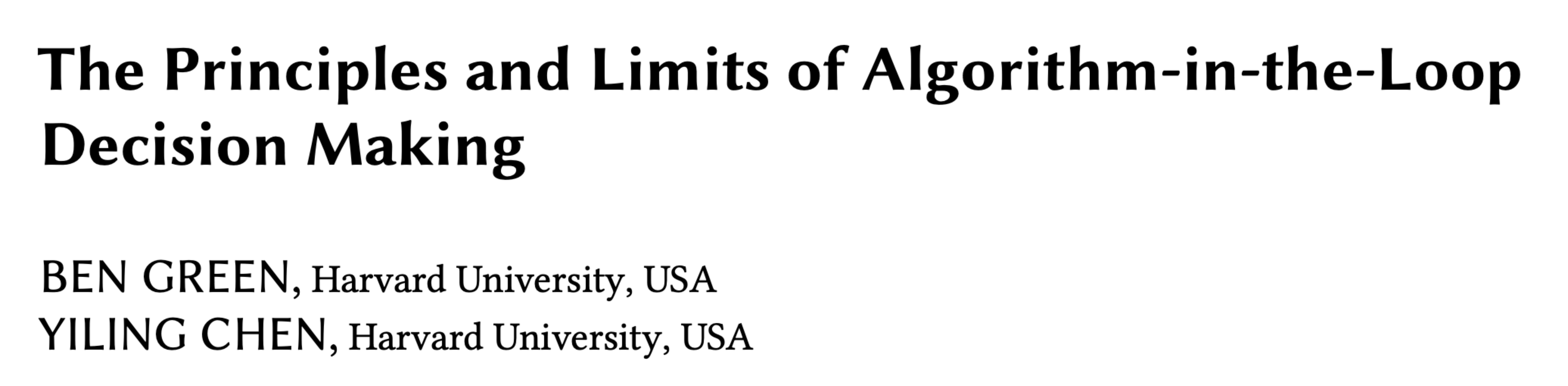


Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments

Ben Green
Harvard University
bgreen@g.harvard.edu

Yiling Chen
Harvard University
yiling@seas.harvard.edu

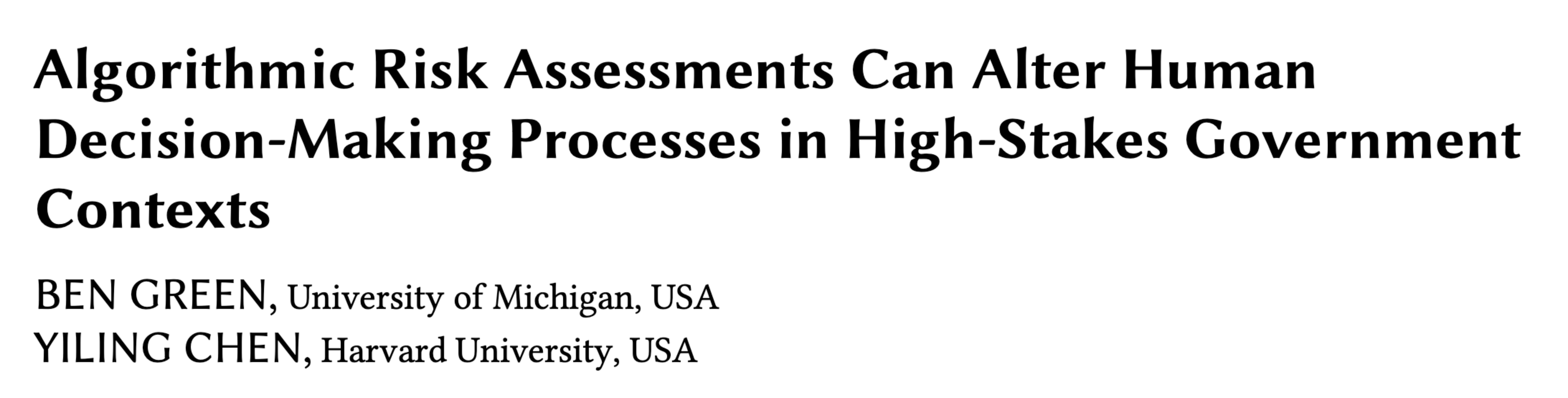
CSCW 2019



The Principles and Limits of Algorithm-in-the-Loop Decision Making

BEN GREEN, Harvard University, USA
YILING CHEN, Harvard University, USA

CSCW 2021



Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts

BEN GREEN, University of Michigan, USA
YILING CHEN, Harvard University, USA

What do human oversight policies propose?

Three Approaches to Human Oversight

Approach 1: Restrict “solely” automated decisions.

GDPR Article 22: “The data subject shall have the right not to be subject to a decision based solely on automated processing.”

Approach 2: Require human discretion.

Northpointe guide to COMPAS: “staff should be encouraged to use their professional judgment and override the computed risk as appropriate.”

Approach 3: Require “meaningful” human input.

Article 29 Data Protection Working Party: “To qualify as human involvement, the controller must ensure that any oversight of the decision is meaningful, rather than just a token gesture.”

Do human oversight policies work?

Two Flaws with Human Oversight Policies

Flaw 1: Human oversight policies are not supported by empirical evidence

Human Discretion Does Not Improve Outcomes

- Across a wide range of domains, automated decision-support systems tend to alter human decision-making in unexpected and detrimental ways.
    - E.g., automation bias (Parasuraman & Manzey, 2010; Skitka et al., 1999).
- People are bad at judging the quality of algorithmic outputs and determining when to override those outputs.
    - Human judgments about how to override an algorithm are typically incorrect and racially biased (Green & Chen, 2019a, 2019b; Grgić-Hlača et al., 2019; Lai & Tan, 2019; Yeomans et al., 2017).
    - Judges often override risk assessments in punitive and racially biased ways (Albright, 2019; Cowgill, 2018; Steinhart, 2006; Stevenson, 2018; Stevenson & Doleac, 2021).

Even “Meaningful” Human Oversight Does Not Improve Outcomes

- The three components of meaningful human oversight are either unlikely to improve decision-making or are very difficult to achieve.
    - People typically override algorithms in detrimental ways.
    - Explanations and transparency do not improve human oversight (Bansal, Wu, et al., 2021; Green & Chen, 2019b; Jacobs et al., 2021; Lai & Tan, 2019; Poursabzi-Sangdeh et al., 2021).
    - People typically defer to automated tools and increase their attention to the factors emphasized by algorithms (Green & Chen, 2021; Parasuraman & Manzey, 2010; Skeem et al., 2019; Starr, 2014).

Two Flaws with Human Oversight Policies

Flaw 1: Human oversight policies are not supported by empirical evidence

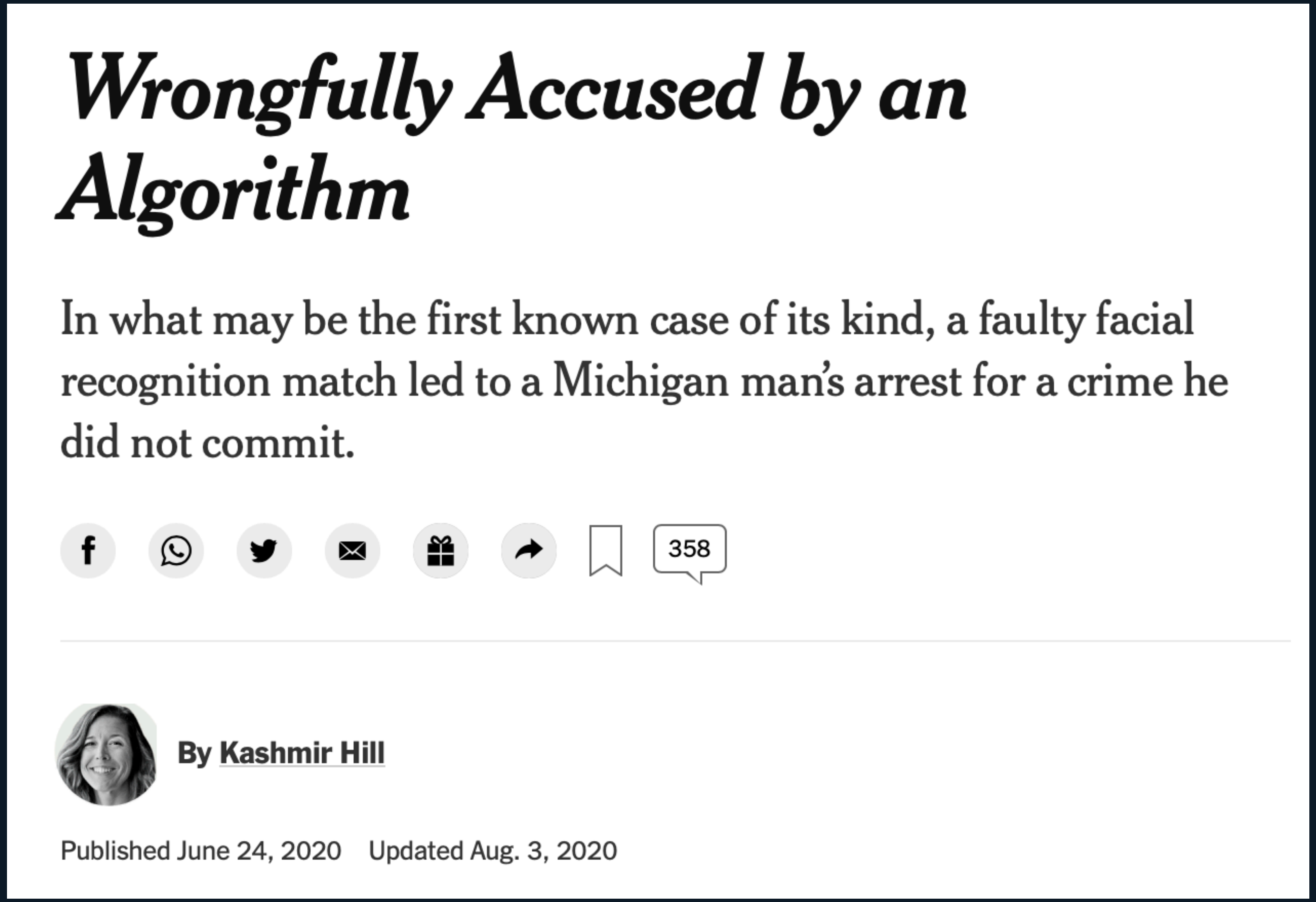
Flaw 2: Human oversight policies legitimize flawed and unaccountable algorithms in government

Human Oversight Provides a False Sense of Security in Adopting Algorithms

- State of Wisconsin v. Loomis:
  - “consideration of COMPAS is permissible; reliance on COMPAS for the sentence imposed is not permissible.”
  - “Just as corrections staff should disregard risk scores that are inconsistent with other factors, we expect that circuit courts will exercise discretion when assessing a COMPAS risk score with respect to each individual defendant.”

Human oversight makes it seem legitimate for governments to use algorithms, but fails to address the underlying concerns about inaccurate and biased predictions.

Human Oversight Diminishes Accountability for Institutional Decision-Makers



Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

By Kashmir Hill

Published June 24, 2020 Updated Aug. 3, 2020

Interview on *60 Minutes*

- Anderson Cooper: What happened in the case of Robert Williams? What went wrong?
- James Craig (Detroit Police Chief): Sloppy, sloppy investigative work. […] That detective was disciplined. And, subsequently, a commanding officer of that command has been de-appointed. But it wasn't facial recognition that failed. What failed was a horrible investigation.

Human oversight positions frontline human operators as the scapegoats for government algorithms, even though the harms of these systems have been structured by more powerful institutional actors.

What’s the alternative?

From Human Oversight to Institutional Oversight

Stage 1: Agency Justification and Evaluation

Stage 2: Democratic Review and Approval

2 Lessons and 2 Open Questions

- We need AI regulation! But the regulation needs to be grounded in empirical evidence about sociotechnical interactions.
- The design and regulation of algorithms needs to consider how people interact with algorithms in practice. What ultimately matters is how algorithms influence human decisions.
  - From “human-in-the-loop” to “algorithm-in-the-loop.”
- What is the political salience and effectiveness of the “human” as an antidote to fears about harmful AI?
- What are principles for determining the appropriate balance/combination of human and algorithmic judgment in decision-making?