Value-Sensitive Design and Responsible Machine Learning

Ethics Lecture for CS 4973-5 (Responsible ML)
Dr. Avijit Ghosh
Fall 2023

Vance Ricks

Associate Teaching Professor of Philosophy and Computer Science

Northeastern University

v.ricks@northeastern.edu

Agenda

- Quick Introduction: machine learning algorithms in the wild
- SOTBF: using a simulation to uncover ethical questions

- Articulating values and identifying stakeholders: value-sensitive design
- From value-sensitive design to values analysis (VAD)
- Three conceptions of “fairness” and “unfairness”
- Treating people as data subjects

- Revisiting SOTBF
- WASTE Assignment overview

- Conclusion: Centering the human in the algorithm

Guiding Assumption 1

"Technology is neither good or bad, nor is it neutral."

Melvin Krantzberg's "First Law of Technology", 1986

How do you interpret this?

Guiding Assumption 2

Unless “no”, “not here”, or “not now” are genuine options, discussions of responsible design and use are purely academic – and not in the good way.

Open access | Research article | First published online February 7, 2022

Resistance and refusal to algorithmic harms: Varieties of ‘knowledge projects’

Maya Indira Ganesh and Emanuel Moss View all authors and affiliations

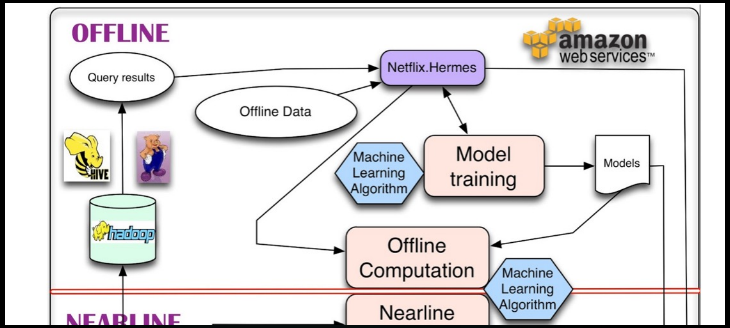
Volume 183, Issue 1 | https://doi.org/10.1177/1329878X221076288

Original Article

Machine Learning Based Computer Aided Diagnosis of Breast Cancer Utilizing Anthropometric and Clinical Features

M.M. Rahman, Y. Ghasemi, E. Suley, Y. Zhou, S. Wang, J. Rogers

How To Design A Spam Filtering System with Machine Learning Algorithm



Some Algorithms In the Wild

More Algorithms In the Wild

Both Zoom and Twitter found themselves under fire this weekend for their respective issues with algorithmic bias. On Zoom, it's an issue with the video conferencing service's virtual backgrounds and on Twitter, it's an issue with the site's photo cropping tool.

It started when Ph.D. student Colin Madland tweeted about a Black faculty member's issues with Zoom. According to Madland, whenever said faculty member would use a virtual background, Zoom would remove his head.

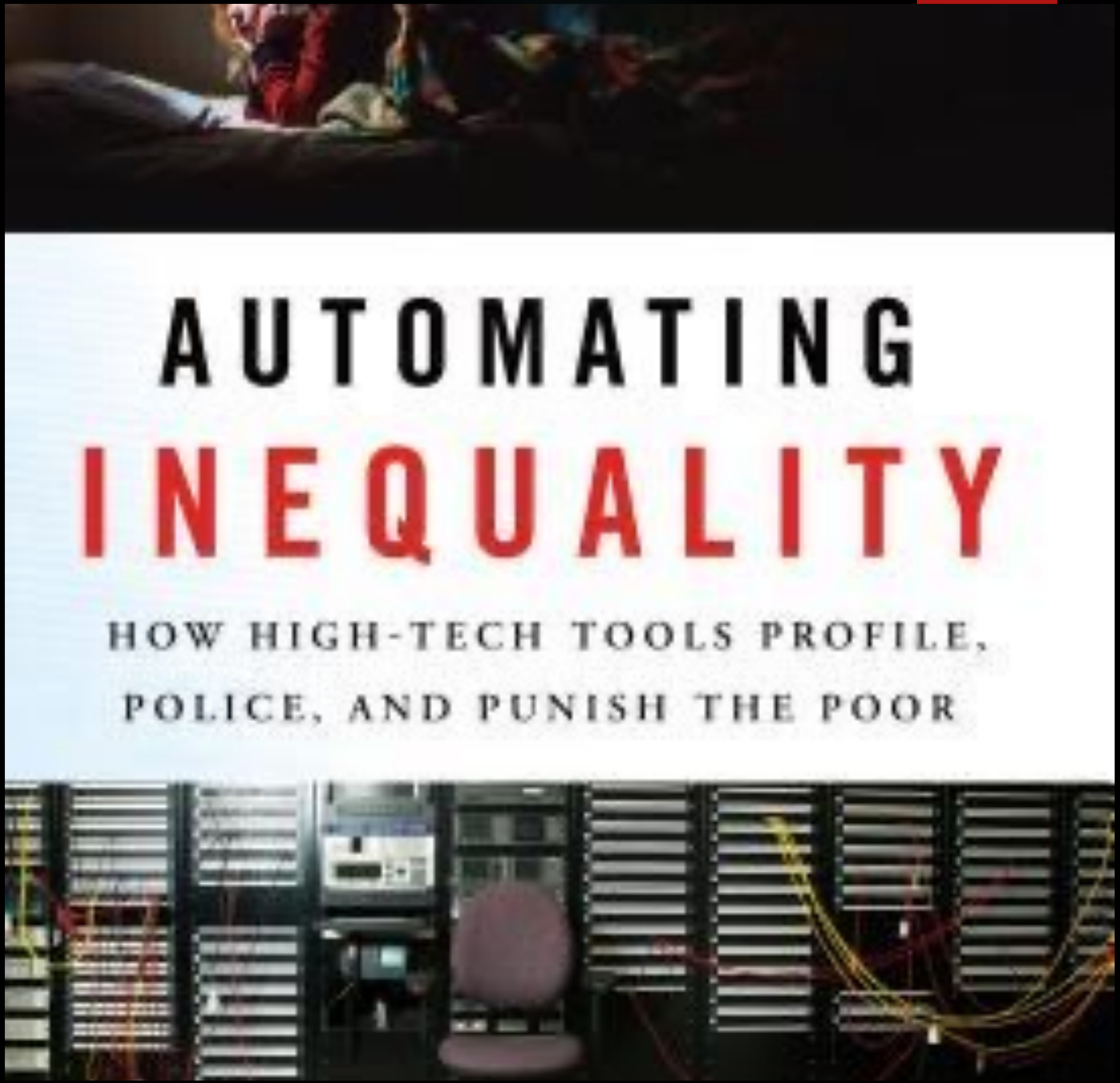
In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

'It's destroyed me completely': Kenyan moderators decry toll of training of AI models

Employees describe the psychological trauma of reading and viewing graphic content, low pay and abrupt dismissals

The Family and Social Services Administration (FSSA) of Indiana provides welfare, food stamps, public health insurance

- goals defined as to **reduce fraud, spending and number of those on welfare**
- prior to automation, FSSA erred on side of providing benefits: False Pos rate = 4.4% False Neg rate = 1.5%
- after automation, erred on opposite side: FP rate = 6.2% FN rate = 12.2%
- when denied, no explanation given for why
- did not use records from previous system, requiring all new applications



Virginia Eubanks, 2019

Yet More Algorithms In the Wild

Eight Months Pregnant and Arrested After False Facial Recognition Match

Porcha Woodruff thought the police who showed up at her door to arrest her for carjacking were joking. She is the first woman known to be wrongfully accused as a result of facial recognition technology.

What Ethics Is, Why It Matters, and How It can Help

What Ethics Isn't (Necessarily)

"It's legal" ≠ "It's ethical"

"It's illegal" ≠ "It's unethical"

What Ethics Isn't (Necessarily)

What Ethics Is

- Ideals, aspirations, standards for how to live well and how to live well *together*
- The uncovering and studying of those ideals and standards
- The clarification, justification, and defense of those ideals and standards
- The living by (or in accordance with) those ideals and standards

Examples of ethical values (NOT an exhaustive list!)

- Accessibility
- Accountability
- Autonomy
- Calm
- Environmental sustainability
- Freedom from bias
- Human welfare
- Identity
- Informed consent
- Ownership / property
- Privacy
- Respect
- Trust

Introducing Value Sensitive Design (VSD)

The case for (the need for) VSD

Technology is the result of human imagination

All technology involves design

All design involves choices among possible options

All choices reflects values

Therefore, all technologies reflect and affect human values

Ignoring values in the design process is irresponsible

Three types of investigation in VSD

Empirical Investigation

- How do **stakeholders** prioritize competing values?
- **Expressed preferences v revealed preferences?**
- **What are the economic incentives in this context?**
- **What are the benefits/costs and their distributions?**

Value Investigation

- What is the **overall goal** of the technology?
- What **values** are at stake?
- Which stakeholders are **legitimately impacted**?
- What value-oriented criteria will be used to gauge project **success**?

Technical Investigation

- How can the tool or system be designed to enable designers to meet their value-oriented goals?
- What effect does **law, policy, and regulation have** on your design?
- Do the technical results **stay within your "red lines"**?

Value Sensitive Design (VSD) in action: the sequence

1. Who are the **stakeholders**? Identify them.
2. What **values** are at stake for those stakeholders? Identify them.
3. Where do there have to be "**tradeoffs**" between some values/interests and other values/interests?
4. Which **core values** need to be given priority, or "**red lines**" need to not be crossed?
5. **Repeat** steps 1 – 4 as you get new information or as circumstances change.
6. Have a clear understanding of a **successful outcome** of this process.

Stakeholders: Whose values / interests are in question?

- **Direct** stakeholders include users, producers, and owners of the technology in question
- **Indirect** stakeholders need to be assessed on a case-by-case basis (people who might not directly interact with the technology in question, but are affected by it nonetheless)
- *Technologies affect more than just those who use them*

What happens when values or interests come into conflict?

Value tradeoffs are needed when:

- multiple values are important;
- they also (seem) hard to achieve at the same time, and so
- a balance must be struck between them

Sometimes this might be different values held by the same party

- e.g., a company that values **security** but also **resource efficiency**
- e.g., should you be a programmer or a nurse?

Sometimes it might be the same value held by different parties

- e.g., **my financial interests** and **the tech company's financial interests**

Can value conflicts be resolved?

- **Assess legitimacy →** are everyone's interests equally legitimate in this context?
- **Respect core values and "red lines" →** are there any values that (almost) cannot be overridden?
- **Promote stronger values →** are there interests or "red lines" that should be prioritized in this context?
- **Understand the social AND technical contexts →** Can some value tensions be revisited or resolved in a different way?

“Success”: Technical v Technological

In CS, we typically think about technical success

- Does the technology function?
- Does it achieve first-order objectives?

Examples:

- Test coverage and bug tracker
- Crash reports
- Benchmarks of speed, prediction accuracy, etc.
- Counts of app installations, user clicks, pages viewed, interaction time, etc.

VSD asks that we think about technological success

- Is the technology beneficial to stakeholders, society, the environment, etc.?
- Is the technology fair or just?

Examples:

- Assessments of quality of life
- Measures of bias
- Reports of bullying, hate speech, etc.
- Carbon footprint

From VSD to VAD

Empirical Investigation

- How do **stakeholders** prioritize competing values?
- **Expressed preferences v revealed preferences?**
- **What are the economic incentives in this context?**
- **What are the benefits/costs and their distributions?**

Value Investigation

- What is the **overall goal** of the technology?
- What **values** are at stake?
- Which stakeholders are **legitimately impacted**?
- What value-oriented criteria will be used to gauge project **success**?

Technical Investigation

- How can the tool or system be designed to enable designers to meet their value-oriented goals?
- What effect does **law, policy, and regulation have** on your design?
- Do the technical results **stay within your "red lines"**?

Preliminary Questions for Small Group (3 – 4 people) Discussions

Instructions:

In your group, take about 5 minutes to discuss and answer the questions below.

Jot down your answers, to report back to the rest of the class.

Question One: What is *fair* treatment, as opposed to *unfair* treatment?

Question Two: Is there a difference between fair *treatment* and a fair *outcome*?

Collected Group Responses – what is *fairness*?

What is fairness in treatment?

What is fairness in outcome(s)?

Three frameworks for thinking about fair treatment

Distributive frameworks

- There's some **good** or **benefit** to be distributed…
- to some **recipients**…
- according to some **distributive principle**…
- that is based on some **underlying values**.

Procedural frameworks

- There's some **good** or **benefit** to be pursued…
- for some **recipients**…
- so we **create a procedure** aimed at achieving that good or benefit.

Interactional frameworks

- There is some decision…
- that will affect some people…
- so we ensure that that decision **respects those people's dignity and interests**.

Three kinds of algorithmic unfairness

Here are three ways that algorithms that automate decision making may fail to treat people fairly:

1) **In their purpose (goals)**: the algorithm is designed to achieve a goal that is *itself* illegitimate, because that goal relies on false assumptions or reinforces attitudes or patterns of unjustified inequality

Two other ways that algorithms that automate decision making may fail to treat people fairly:

2) **In their data collection practices (training data)**: the algorithm is not as *accurate* as it could be because of poorly chosen target variables, underlying bias reproduced in training examples, unrepresentative samples, or coarse features

3) **In their distribution of burdens of error (outcomes)**: the data and algorithm are as good as possible, but the algorithm imposes greater burdens of error on some stakeholders than others, often in ways that reinforce existing patterns of inequality in society

1. **In their purposes (bad or flawed goal)**

Ones based on empirically false assumptions

Ones with a foreseeable high risk of making already-vulnerable groups even more vulnerable

Example of Empirically False Assumptions

POLICY

AI 'EMOTION RECOGNITION' CAN'T BE TRUSTED

The belief that facial expressions reliably correspond to emotions is unfounded, says a new review of the field

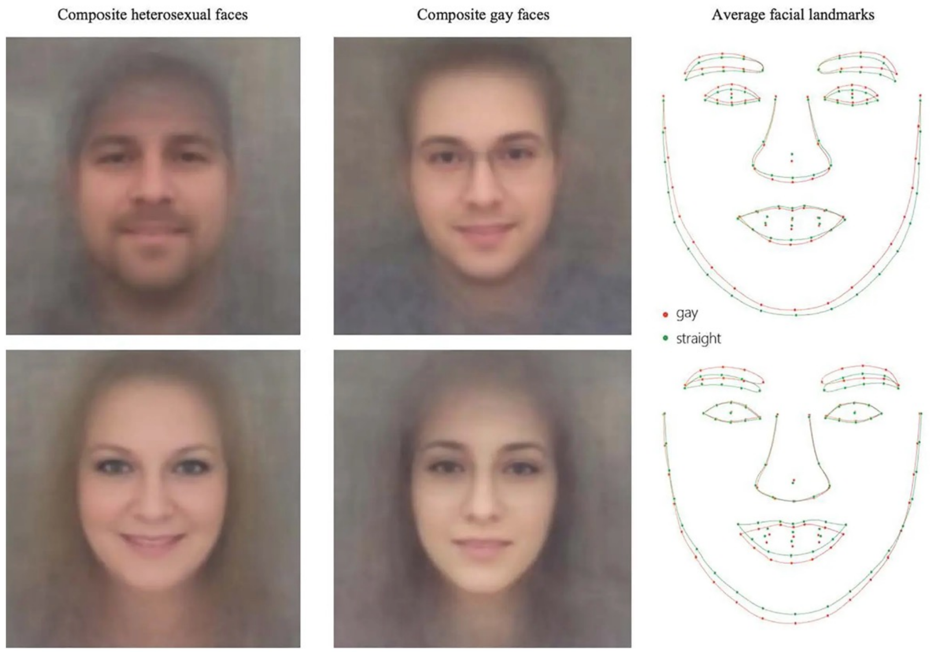
By James Vincent | Jul 25, 2019, 11:55am EDT

But the belief that we can easily infer how people feel based on how they look is controversial, and a significant new review of the research suggests there's no firm scientific justification for it.

"Companies can say whatever they want, but the data are clear," Lisa Feldman Barrett, a professor of psychology at Northeastern University and one of the review's five authors, tells *The Verge*. "They can detect a scowl, but that's not the same thing as detecting anger."

Example of increasing vulnerability

Why Stanford Researchers Tried to Create a ‘Gaydar’ Machine



In one section of the study, the authors presented what they called “composite heterosexual faces,” left, and “composite gay faces,” right, built by averaging landmark locations of the faces classified as most and least likely to be gay. This did not sit well with some who suggested that this is just the latest example of physiognomy, rationalized by deep neural networks. Michal Kosinski and Yilun Wang

New York Times, Oct 9, 2017

For all these reasons, there's a growing recognition among scholars and advocates that some biased AI systems should not be "fixed," but abandoned. As co-author Meredith Whittaker said, "We need to look beyond technical fixes for social problems. We need to ask: Who has power? Who is harmed? Who benefits? And ultimately, who gets to decide how these tools are built and which purposes they serve?"

"It's not biased" ≠ "It's morally harmless"

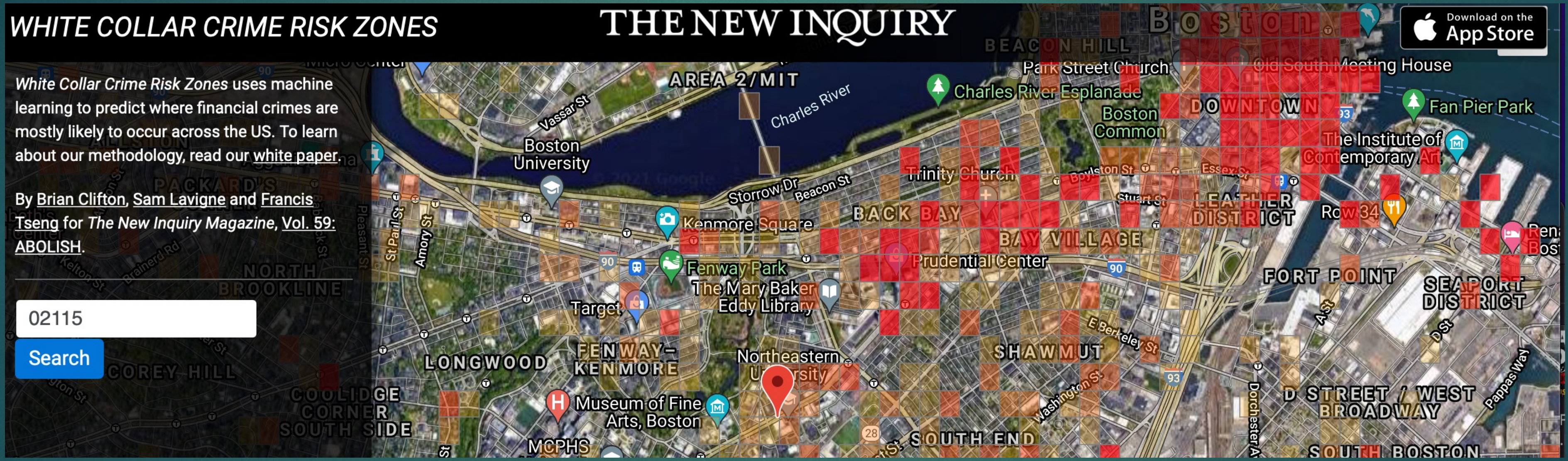
From *Vox*, "Some AI just shouldn't exist", 19 April 2019

2. **In Data Collection Practices**

Sources of bad or biased training data

a. When defining target variables and in class labels
b. When assembling the training data set, resulting in an unrepresentative sample
c. When selecting relevant features
d. Intentional bias: masking, redlining, etc.
e. Treatment of the data sources and labelers

How are the categories defined? (e.g., "crime")



How are the data subjects and labelers treated?

Intellectual property concerns

Labor rights concerns

3. In Distribution of Burdens of Algorithmic Error (in decisions or outcomes)

Treating People as Data Subjects

The tension:

“constructing the human *as a data point for machine training and optimization* rather than *as a person who should be justly, equitably, and sensitively treated*”

(Chancellor et al., p 2)

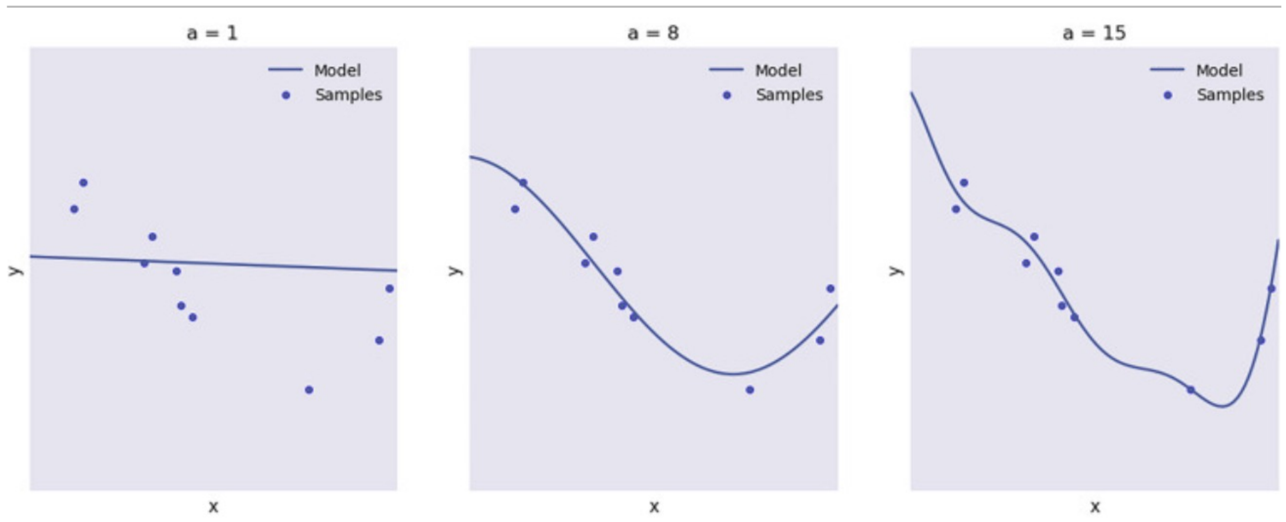
Summing Up: some ways to address unfairness in algorithms

How do we avoid

(creating or relying on algorithms that end up)

treating people unfairly?

Zeroth, remember that the model itself, not just the data, could be a problem



Download : Download high-res image (156KB)

Download : Download full-size image

Figure 1. Our model choices express a preference for model behavior. An example most students of machine learning will recognize is the plot between the degrees of a polynomial (a) and the degree of overfitting.

From Hooker, "Moving Beyond 'Algorithmic Bias Is A Data Problem"

First, pay careful attention to how data is collected and classified

- In how the collectors and labelers are treated
- When defining target variables and in class labels
- When assembling the training data set, resulting in an unrepresentative sample
- When selecting relevant features
- Watch out for *intentional* bias: masking, redlining, etc.

Second, make explicit ethical decisions about how to distribute the unfairness

- Even if the algorithm is "perfectly accurate", there might still be some unfairness **because of the social context in which it is used**
- To distribute the risks of error more fairly, you should at a minimum bring in all stakeholders
- Consider whether an algorithm should be used **at all** in this domain (e.g., perhaps any foreseeable algorithmic error in criminal justice contexts sentences is ethically intolerable?)

Small-group activity

Apply VSD/VAD analyses to the following case:

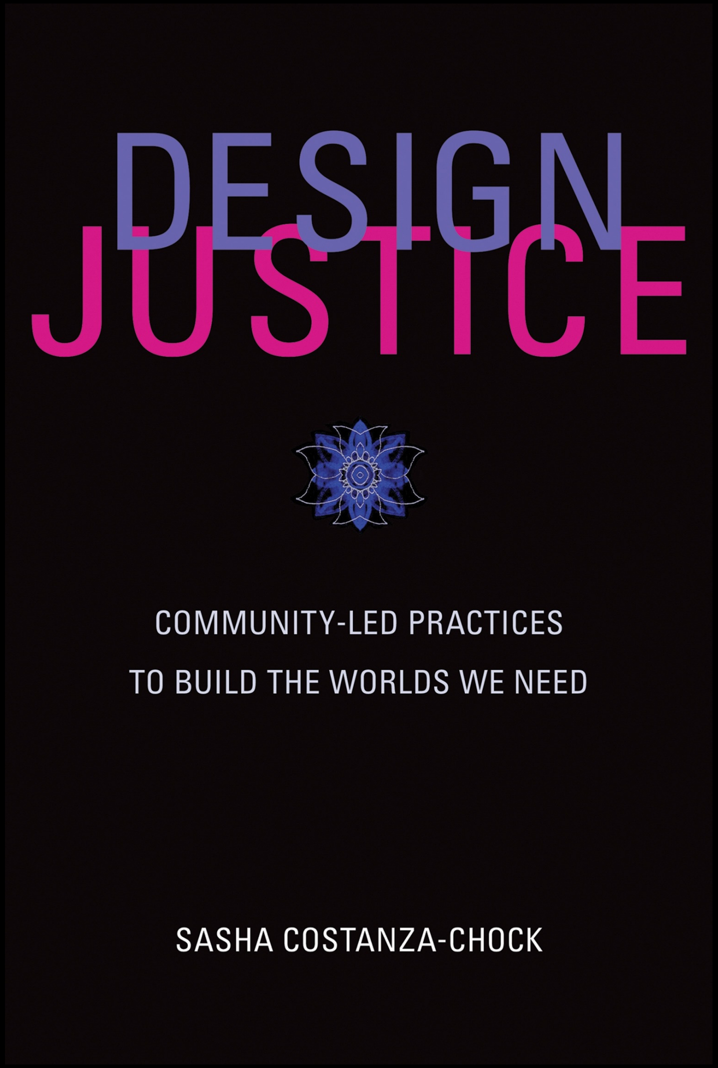
(Made-up example)

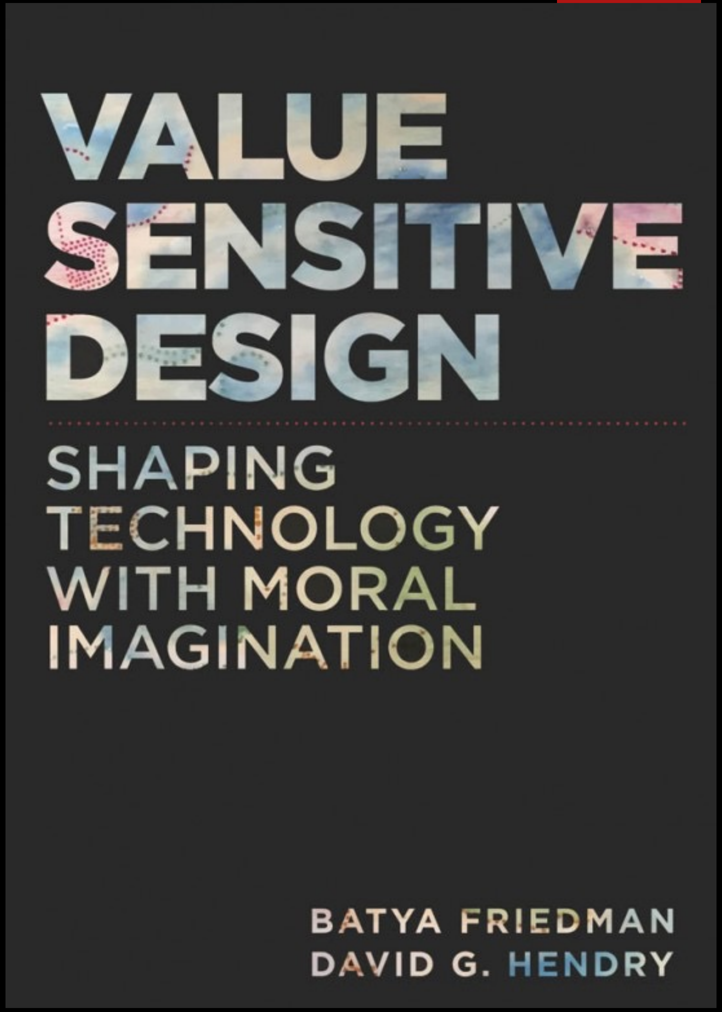
A city ordinance written by legislators in a medium-sized USA city with an older, dense downtown that is surrounded by suburbs.

We must, therefore, make careful, explicit choices as to how and where to distribute the burdens of error in the algorithms we build.

This should be done at both the **law and policy** level, and at the **design** level, which is where value-sensitive design – an approach that emphasizes stakeholder interests and values – attempts to intervene.

We should also ask ***whether an algorithm should be used at all*** for the task at hand.





Thank you!

Some review questions:

- What does it mean to treat people fairly?
- What are three main ways that algorithms that automate decision-making might treat people unfairly?
- Why are there necessarily trade-offs between these measures of fairness in algorithmic design?
- How should we deal with such trade-offs? What should we do about them?