

CS 4530

Fundamentals of Software Engineering

Module 18: Engineering Software for Equity

Jonathan Bell, Adeel Bhutta, Jan Vitek, Mitch Wand
Khoury College of Computer Sciences

Learning Goals

By the end of this lesson, you should be able to...

- ◉ Illustrate how software can cause inadvertent harm or amplify inequities
- ◉ Explain the role of software engineers in avoiding such harms

Equity and Software

*As new as software engineering is,
we're newer still at understanding its impact on underrepresented
people and diverse societies*

*We must recognize imbalance of power between those who make
development decisions that impact the world
and those who simply must accept and live with those decisions that
sometimes disadvantage already marginalized communities globally*

Recognize inequities in your software



One mark of an exceptional engineer is the ability to understand how products can advantage and disadvantage different groups of human beings

Engineers are expected to have technical aptitude, but they should also have the discernment to know when to build something and when not to

Demma Rodriguez
Head of Equity Engineering
Google

Recognize inequities in your software

- ◉ Good engineers understand how products can be weaponized to create harms in certain groups
- ◉ Microsoft failed with a chatbot that picked up the behavior people used...
- ◉ ...they taught Tay to use offensive and racist language attacking jews



Recognize inequities in your software

- ◉ Good engineers understand how products can be weaponized...
- ◉ Amazon failed with their AI hiring software...
- ◉ ...it used 10 years of resumes to learn who should be hired
- ◉ ...it learned to automatically reject the resumes of women



Algorithmic sentencing discriminates

The COMPAS sentencing tool discriminates against black defendants

| | ALL | WHITE DEFENDANTS | BLACK DEFENDANTS |
|---|-----|------------------|------------------|
| Labeled High Risk , But Didn't Re-Offend | 32% | 23% | 44% |
| Labeled Low Risk , Yet Did Re-Offend | 37% | 47% | 28% |

Algorithmic bias discriminates

...against the poorest of us

THE WALL STREET JOURNAL.

Websites Vary Prices, Deals Based on Users' Information

Getting Different Deals Online

A Journal examination found online retailers adjusted prices by a shopper's location, among other factors

Staples.com

SnapSafe Titan safe

HIGHER PRICE
\$1,199.99
DISCOUNT PRICE
\$1,099.99

DIFFERENCE:
9.1%

Homedepot.com

A 250-foot spool of electrical wiring



Six pricing groups, including:

\$70.80 in Ashtabula, Ohio
\$72.45 in Erie, Pa.
\$77.87 in Monticello, N.Y.

Rosettastone.com



...for buying multiple levels of German lessons, when test-shopping from the U.S. or Canada. But not from the U.K. or Argentina.

Photos: l to r: SnapSafe; Home Depot; Rosetta Stone Source: WSJ testing

The Wall Street Journal

FairTest: Discovering Unwarranted Associations in Data-Driven Applications*

Florian Tramèr¹, Vaggelis Atlidakis², Roxana Geambasu², Daniel Hsu²,
Jean-Pierre Hubaux³, Mathias Humbert⁴, Ari Juels⁵, Huang Lin³

¹Stanford, ²Columbia University, ³EPFL, ⁴Saarland University, ⁵Cornell Tech, Jacobs Institute

Abstract—In a world where traditional notions of privacy are increasingly challenged by the myriad companies that collect and analyze our data, it is important that decision-making entities are held accountable for unfair treatments arising from irresponsible data usage. Unfortunately, a lack of appropriate methodologies and tools means that even identifying unfair or discriminatory effects can be a challenge in practice.

We introduce the *unwarranted associations (UA) framework*, a principled methodology for the discovery of unfair, discriminatory, or offensive user treatment in data-driven applications. The UA framework unifies and rationalizes a number of prior attempts at formalizing algorithmic fairness. It uniquely combines multiple investigative primitives and fairness metrics with broad applicability, granular exploration of unfair treatment in user subgroups, and incorporation of natural notions of utility that may account for observed disparities.

We instantiate the UA framework in *FairTest*, the first comprehensive tool that helps developers check data-driven applications for unfair user treatment. It enables scalable and statistically rigorous investigation of associations between application outcomes (such as prices or premiums) and sensitive user attributes (such as race or gender). Furthermore, *FairTest* provides *debugging capabilities* that let programmers rule out

decision-making can have unintended and harmful consequences, such as unfair or discriminatory treatment of users.

In this paper, we deal with the latter challenge. Despite the personal and societal benefits of today's data-driven world, we argue that companies that collect and use our data have a responsibility to ensure equitable user treatment. Indeed, European and U.S. regulators, as well as various policy and legal scholars, have recently called for increased *algorithmic accountability*, and in particular for decision-making tools to be audited and “tested for fairness” [1], [2].

There have been many recent reports of unfair or discriminatory effects in data-driven applications, mostly qualified as unintended consequences of data heuristics or overlooked bugs. For example, Google's image tagger was found to associate racially offensive labels with images of black people [3]; the developers called the situation a bug and promised to remedy it as soon as possible. In another case [4], *Wall Street Journal* investigators showed that Staples' online pricing algorithm discriminated against lower-income people. They referred to the situation as an “unintended consequence” of Staples's seemingly rational decision to adjust online prices based on user proximity to competitors' stores. This led to higher prices for low-income customers who generally live farther from these stores

Training AI systems impacts climate

TheRegister®

≡

{* AI + ML *}

AI me to the Moon... Carbon footprint for 'training GPT-3' same as driving to our natural satellite and back

Get ready for Energy Star stickers on your robo-butlers, maybe?

Katyanna Quach

Wed 4 Nov 2020 // 07:59 UTC

SHARE

Training OpenAI's giant GPT-3 text-generating model is akin to driving a car to the Moon and back, computer scientists reckon.

More specifically, they estimated teaching the **neural super-network** in a Microsoft data center using Nvidia GPUs required roughly 190,000 kWh, which using the average carbon intensity of America would have produced 85,000 kg of CO₂ equivalents, the same amount produced by a new car in Europe driving 700,000 km, or 435,000 miles, which is about twice the distance between Earth and the Moon, some 480,000 miles.

Phew.

https://www.theregister.com/2020/11/04/gpt3_carbon_footprint_estimate/

| Consumption | CO ₂ e (lbs) |
|---------------------------------|-------------------------|
| Air travel, 1 passenger, NY↔SF | 1984 |
| Human life, avg, 1 year | 11,023 |
| American life, avg, 1 year | 36,156 |
| Car, avg incl. fuel, 1 lifetime | 126,000 |
| Training one model (GPU) | |
| NLP pipeline (parsing, SRL) | 39 |
| w/ tuning & experimentation | 78,468 |
| Transformer (big) | 192 |
| w/ neural architecture search | 626,155 |

[“Energy and Policy Considerations for Deep Learning in NLP”](#)
by Strubell et al in ACL19

9

UIs discriminate against differently-abled

Inclusivity and Accessibility: Domino’s Pizza LLC v. Robles

Domino’s Would Rather Go to the Supreme Court Than Make Its Website Accessible to the Blind

Rather than developing technology to support users with disabilities, the pizza chain is taking its fight to the top

by Brenna Houck | @EaterDetroit | Jul 25, 2019, 6:00pm EDT

   SHARE

[“Domino’s Would Rather Go to the Supreme Court Than Make Its Website Accessible to the Blind” by Brenna Houck, Eater Detroit](#)



| | |
|-------------|--|
| Jul 15 2019 | Brief amicus curiae of Washington Legal Foundation filed. |
| Jul 15 2019 | Brief amici curiae of Retail Litigation Center, Inc., et al. filed. |
| Jul 15 2019 | Brief amicus curiae of Cato Institute filed. |
| Jul 15 2019 | Brief amicus curiae of Restaurant Law Center filed. |
| Jul 15 2019 | Brief amici curiae of Chamber of Commerce of the United States America, et al. filed. |

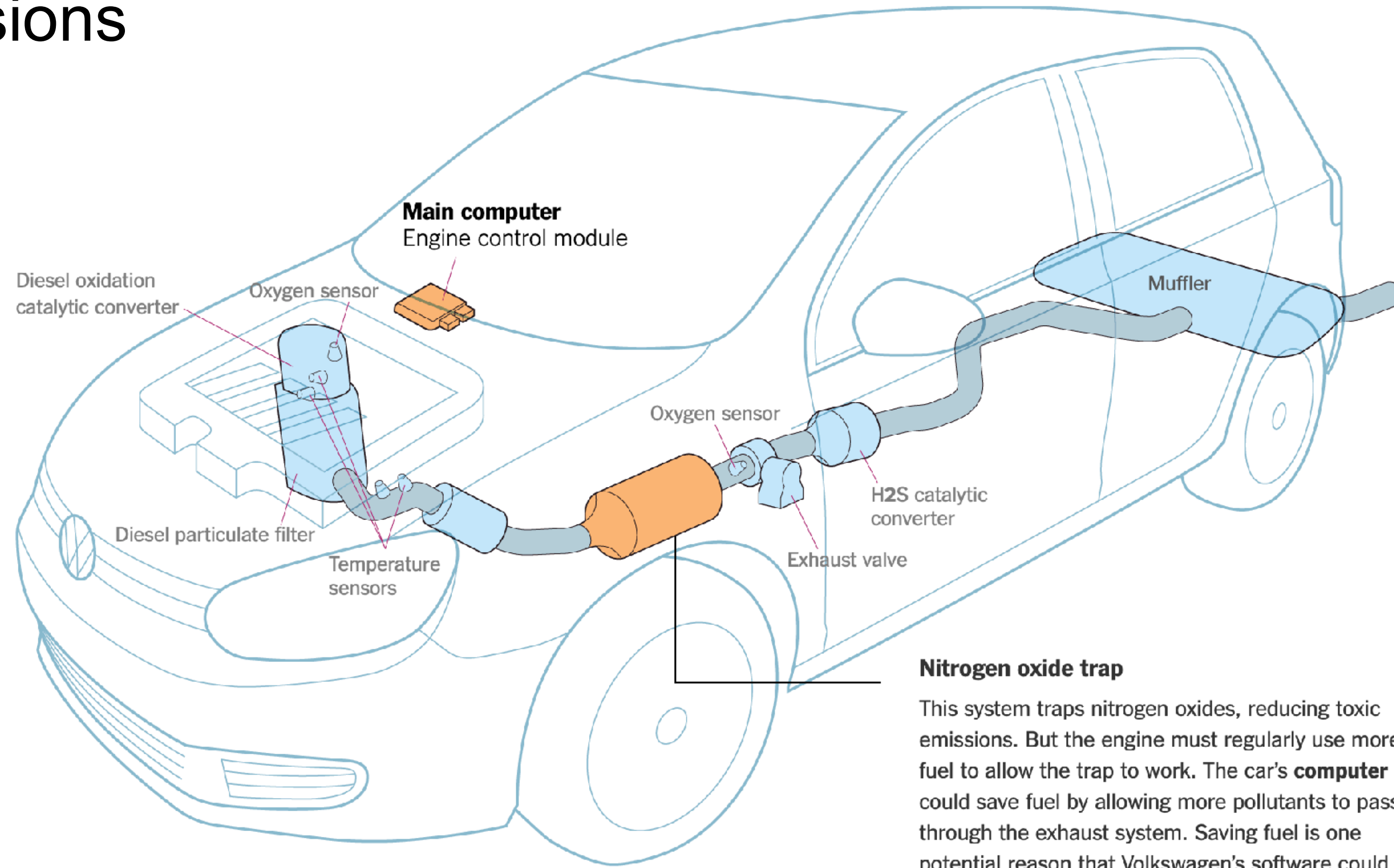
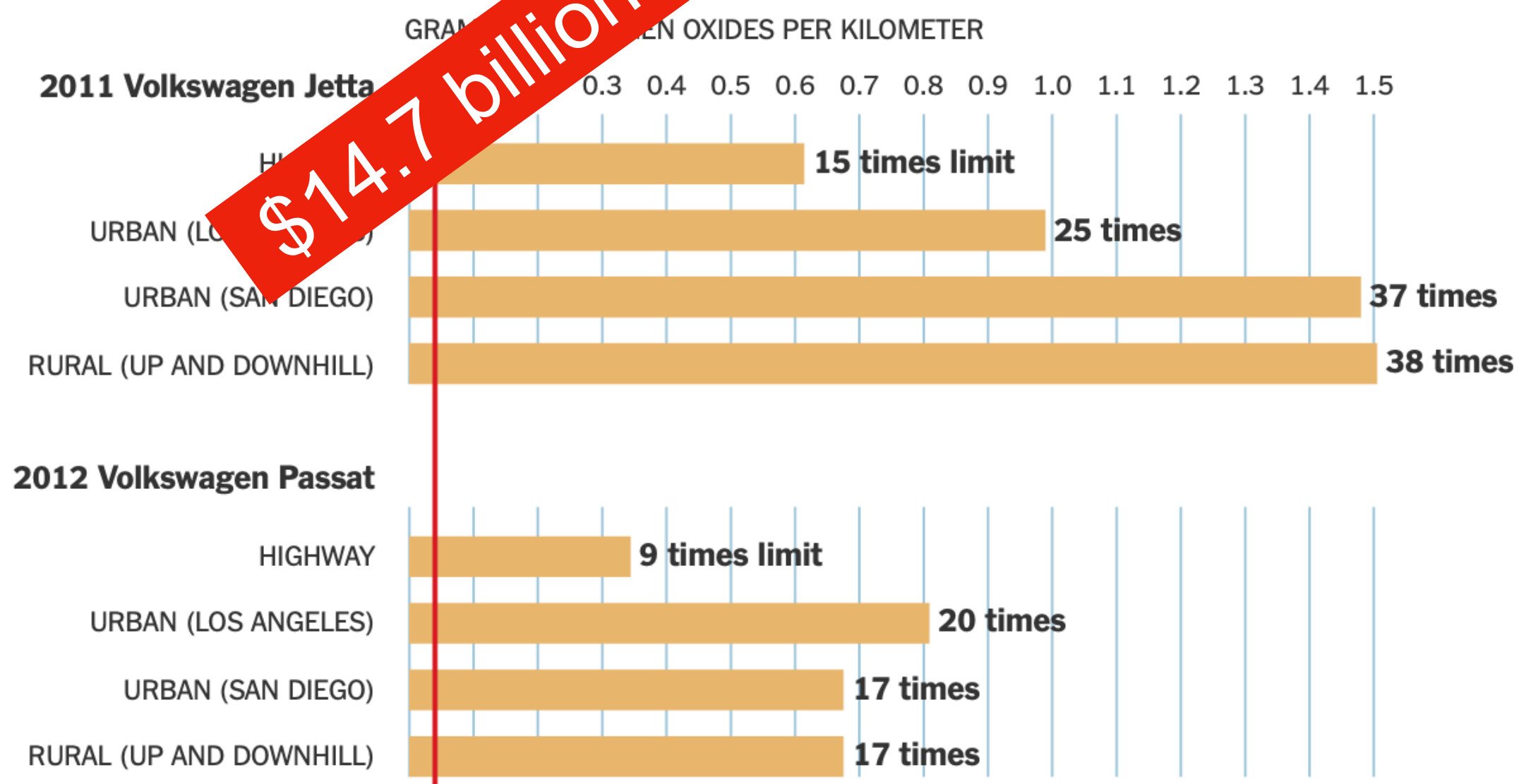
Software evades regulation

Example: Volkswagen diesel emissions

The Emissions Tests That Led to the Discovery of VW's Cheating

The on-road testing in May 2014 that led the California Air Resources Board to investigate Volkswagen was conducted by researchers at West Virginia University. They tested emissions from two VW models equipped with the 2-liter turbocharged 4-cylinder diesel engine. The researchers found that when tested on the road, some cars emitted almost **40 times** the permitted level of nitrogen oxides.

Average emissions of nitrogen oxides during on-road testing



Nitrogen oxide trap
This system traps nitrogen oxides, reducing toxic emissions. But the engine must regularly use more fuel to allow the trap to work. The car's **computer** could save fuel by allowing more pollutants to pass through the exhaust system. Saving fuel is one potential reason that Volkswagen's software could have been altered to make cars pollute more, according to researchers at the International Council on Clean Transportation.

Illustration by Guilbert Gates | Source: Volkswagen, The International Council on Clean Transportation

Bias is the Default

Example: Google Photos auto-tagging



THE WALL STREET JOURNAL.



DIGITS

Google Mistakenly Tags Black People as ‘Gorillas,’ Showing Limits of Algorithms

By [Alistair Barr](#)

Updated July 1, 2015 3:41 pm ET

SHARE TEXT

Google is a leader in artificial intelligence and machine learning. But the company’s computers still have a lot to learn, judging by a major blunder by its Photos app this week.

The app tagged two black people as “Gorillas,” according to Jacky Alciné, a Web developer who spotted the error and tweeted a photo of it.

“Google Photos, y’all f**ked up. My friend’s not a gorilla,” [he wrote on Twitter](#).

<https://www.wsj.com/articles/BL-DGB-42522>

<https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>



WIRED

BACKCHANNEL BUSINESS CULTURE GEAR IDEAS MORE SIGN IN

TOM SIMONITE

BUSINESS 01.11.2018 07:00 AM

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

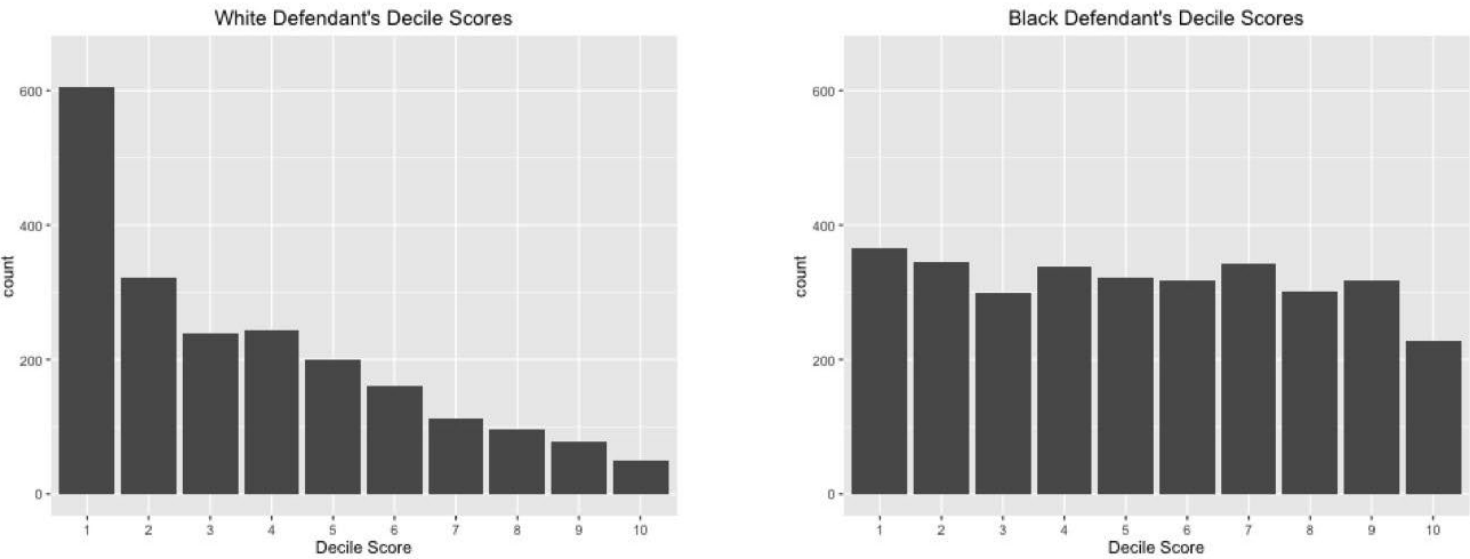


Reflecting on these examples

Personal philosophies and business cases

Algorithmic Bias: COMPAS Sentencing Tool

| | ALL DEFENDANTS | WHITE DEFENDANTS | BLACK DEFENDANTS |
|---|----------------|------------------|------------------|
| Labeled Higher Risk, But Didn't Re-Offend | 32.4% | 23.5% | 44.9% |
| Labeled Lower Risk, Yet Did Re-Offend | 37.4% | 47.7% | 28.0% |



Analysis of Broward County, FL data: “How We Analyzed the COMPAS Recidivism Algorithm” by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin

Algorithmic Bias: Price Discrimination

THE WALL STREET JOURNAL.

Websites Vary Prices, Deals Based on Users' Information

Getting Different Deals Online

A Journal examination found online retailers adjusted prices by a shopper's location, among other factors

Staples.com: SnapSafe Titan safe. Higher price: \$1,199.99. Discount price: \$1,099.99. Difference: 9.1%.

HomeDepot.com: A 250-foot spool of electrical wiring. Six pricing groups, including: \$70.80 in Ashtabula, Ohio; \$72.45 in Erie, Pa.; \$77.87 in Monticello, NY.

Rosettastone.com: A 20% discount.

By Jennifer Valentino-DeVries, Jeremy Singer-Vine and Ashkan Soltani
December 24, 2012
<https://www.wsj.com/articles/SB1000142412788732377204578189391813881534>

2017 IEEE European Symposium on Security and Privacy

FairTest: Discovering Unwarranted Associations in Data-Driven Applications*

Florian Tramèr¹, Vaggelis Athidakis², Roxana Geambasu³, Daniel Hsu⁴, Jean-Pierre Hubaux¹, Mathias Humbert¹, Avi Juels⁵, Huang Lin⁶

¹Stanford, ²Columbia University, ³EPFL, ⁴Saarland University, ⁵Cornell Tech, Jacobs Institute

Abstract—In a world where traditional notions of privacy are increasingly challenged by the myriad companies that collect and analyze our data, it is important that decision-making entities are held accountable for unfair treatments arising from irresponsible data usage. Unfortunately, a lack of appropriate methodologies and tools means that even identifying unfair or discriminatory effects can be a challenge in practice.

We introduce the *unwarranted associations (UIA) framework*, a principled methodology for the discovery of unfair, discriminatory, or offensive user treatment in data-driven applications. The UIA framework unifies and rationalizes a number of prior attempts at formalizing algorithmic fairness. It uniquely combines multiple investigative primitives and fairness metrics with broad applicability, granular exploration of unfair treatment in user subgroups, and incorporation of natural notions of utility that may account for observed disparities.

We instantiate the UIA framework in *FairTest*, the first comprehensive tool that helps developers check data-driven applications for unfair user treatment. It enables scalable and statistically rigorous investigation of associations between application outcomes (such as prices or premiums) and sensitive user attributes (such as race or gender). Furthermore, *FairTest* provides *debugging capabilities* that let programmers rule out potential confounders for observed unfair effects.

We report on use of *FairTest* to investigate and in some cases address disparate impact, offensive labeling, and uneven rates of algorithmic error in four data-driven applications. As examples, our results reveal subtle biases against older populations in the distribution of error in a predictive health application and offensive racial labeling in an image tagger.

We argue that such algorithmic biases are new kinds of bugs, specific to modern, data-driven applications, that programmers should proactively check for, debug, and fix with the same rigor as they apply to other security and privacy bugs. Such bugs can offend and even harm users, and cause programmers and businesses embarrassment, mistrust, and potentially loss of revenue. They may also be symptoms of a malfunction of a data-driven algorithm, such as a ML algorithm exhibiting poor accuracy for minority groups that are underrepresented in its training set [3].

We refer to such bugs generally as *unwarranted associations*. Understanding and identifying unwarranted associations is an important step towards holding automated decision-making entities accountable for unfair practices, thus also providing incentive for the adoption of corrective measures [1], [2], [6], [7].

*Work done while the first author was at EPFL.

© 2017, Florian Tramèr. Under license to IEEE. DOI: 10.1109/EuroSP.2017.29

401

© copyright

Inclusivity and Accessibility: Domino’s Pizza LLC v. Robles

Domino’s Would Rather Go to the Supreme Court Than Make Its Website Accessible to the Blind

Rather than developing technology to support users with disabilities, the pizza chain is taking its fight to the top

by Brenna Houck | @EaterDetroit | Jul 25, 2019, 6:00pm EDT

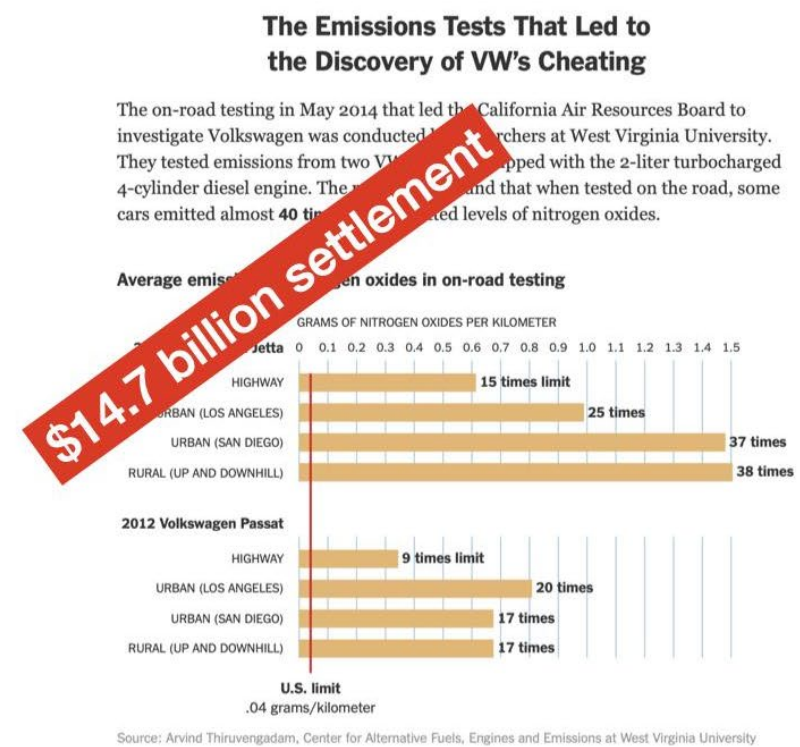
f SHARE



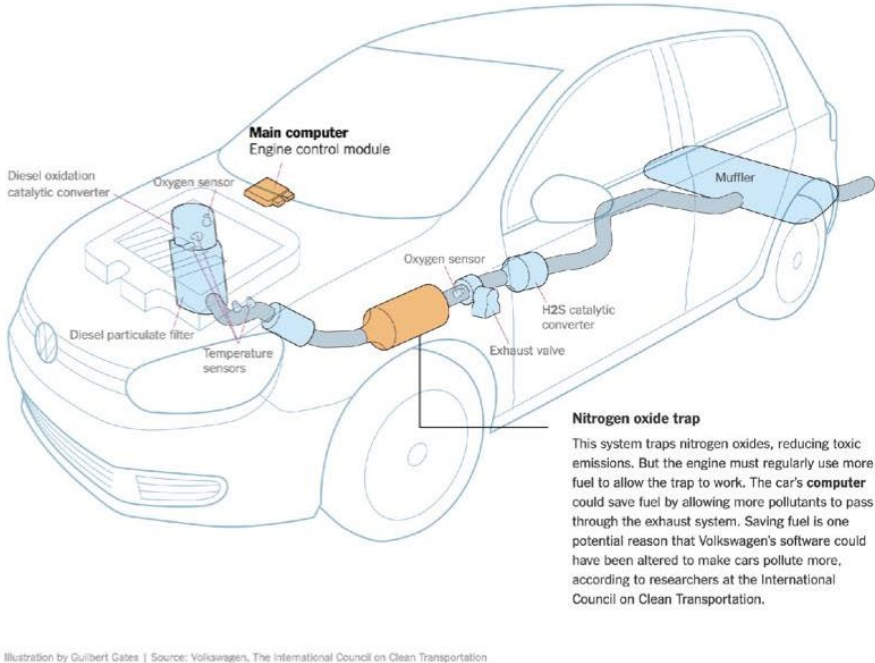
| | |
|-------------|---|
| Jul 15 2019 | Brief amicus curiae of Washington Legal Foundation filed. |
| Jul 15 2019 | Brief amici curiae of Retail Litigation Center, Inc., et al. filed. |
| Jul 15 2019 | Brief amicus curiae of Cato Institute filed. |
| Jul 15 2019 | Brief amicus curiae of Restaurant Law Center filed. |
| Jul 15 2019 | Brief amici curiae of Chamber of Commerce of the United States of America, et al. filed. |

“Domino’s Would Rather Go to the Supreme Court Than Make Its Website Accessible to the Blind” by Brenna Houck, Eater Detroit

Evading regulation: Volkswagen



“How Volkswagen’s ‘Defeat Devices’ Worked” By Guilbert Gates, Jack Ewing, Karl Russell and Derek Watkins



More than *don't be evil*

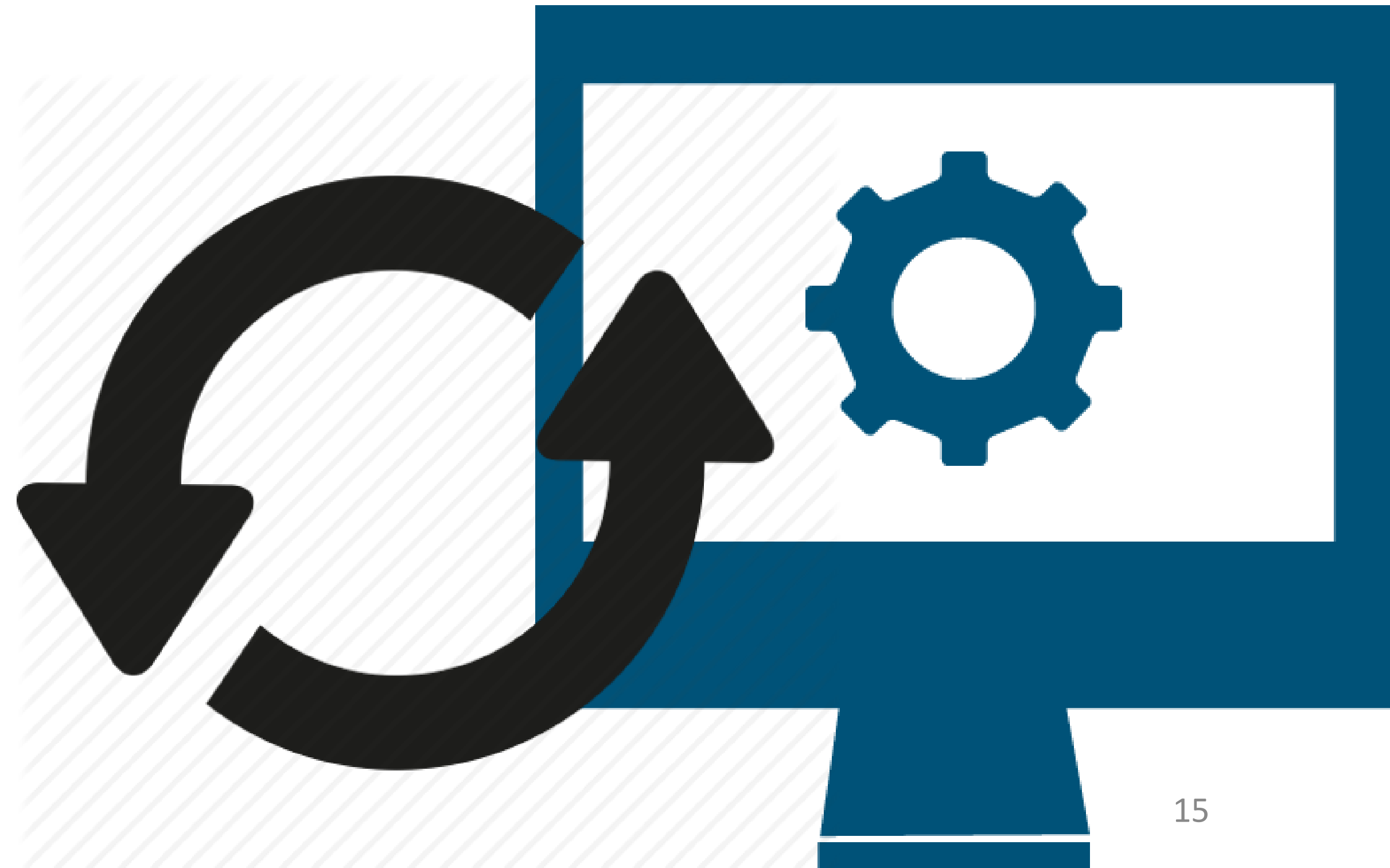
Engineering equitable software requires conscious effort

- ⦿ How do we determine what “the right thing” is?
- ⦿ How do we convince our investors/managers to take this action?

How to mitigate harms in software?

What are you trying to solve?

- ◉ For every software you create, include a wide range of people to use it
- ◉ Including more people helps detect biases and harms
- ◉ Iterate your software throughout its entire life cycle.



How to write software for people that mitigates harm



Microsoft

1

INITIALLY

**Make clear what
the system can do**

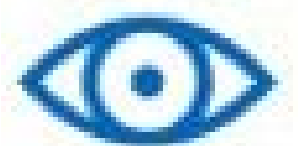
Help the users understand what
the AI system is capable of doing.

2

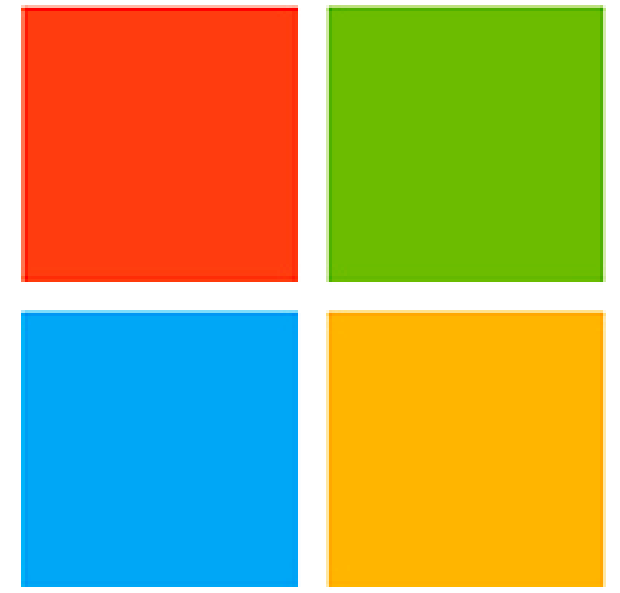
INITIALLY

**Make clear how
well the system can
do what it can do.**

Help the user understand how
often the AI system may make
mistakes.



INITIALLY



Microsoft

3

DURING INTERACTION

Time services based on context.

Time when to act or interrupt based on the user's current task and environment.

4

DURING INTERACTION

Show contextually relevant information.

Display information relevant to the users' current task and environment.

5

DURING INTERACTION

Match relevant social norms.

Ensure the experience is delivered in a way that users would expect, given their social and cultural context.

6

DURING INTERACTION

Mitigate social biases.

Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.

 DURING INTERACTION



Microsoft

7

WHEN WRONG

Support efficient invocation.

Make it easy to invoke or request the AI system's services when needed.

8

WHEN WRONG

Support efficient dismissal.

Make it easy to dismiss or ignore undesired system services.

9

WHEN WRONG

Support efficient correction.

Make it easy to edit, refine, or recover when the AI system is wrong.

10

WHEN WRONG

Scope services when in doubt.

Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.

11

WHEN WRONG

Make clear why the system did what it did.

Enable the user to access an explanation of why the AI system behaved as it did.

 WHEN WRONG



Microsoft

12
OVER TIME

Remember recent interactions.

Maintain short-term memory and allow the user to make efficient references to that memory.

13
OVER TIME

Learn from user behavior.

Personalize the user's experience by learning from their actions over time.

14
OVER TIME

Update and adapt cautiously.

Limit disruptive changes when updating and adapting the AI system's behaviors.

15
OVER TIME

Encourage granular feedback.

Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.

16
OVER TIME

Convey the consequences of user actions.

Immediately update or convey how user actions will impact future behaviors of the AI system.

🕒 OVER TIME

17
OVER TIME

Provide global controls.

Allow the user to globally customize what the AI system monitors and how it behaves.

18
OVER TIME

Notify users about changes.

Inform the user when the AI system adds or updates its capabilities.

Learning Goals

You should now be able to...

- ◉ Suggest how software can cause inadvertent harm or amplify inequities
- ◉ Explain why software engineers have a role to play in avoiding such harms

Exercise

Team up and propose actionable ideas to re-design Amazon's Hiring Software

