CS 4530 & CS 5500 Software Engineering

Lecture 11.1: Engineering Equitable Software

Jonathan Bell, John Boyland, Mitch Wand
Khoury College of Computer Sciences

Learning Objectives for this Lesson

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

- Explain that just because you *can* build some software does not mean that you should
- Provide examples of situations where software causes (inadvertent) harm

Engineering Equitable 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

Quote: “Software Engineering at Google: Lessons Learned from Programming Over Time,” Wright, Winters and Manshreck, 2020 (O’Reilly)

More than “don’t be evil”

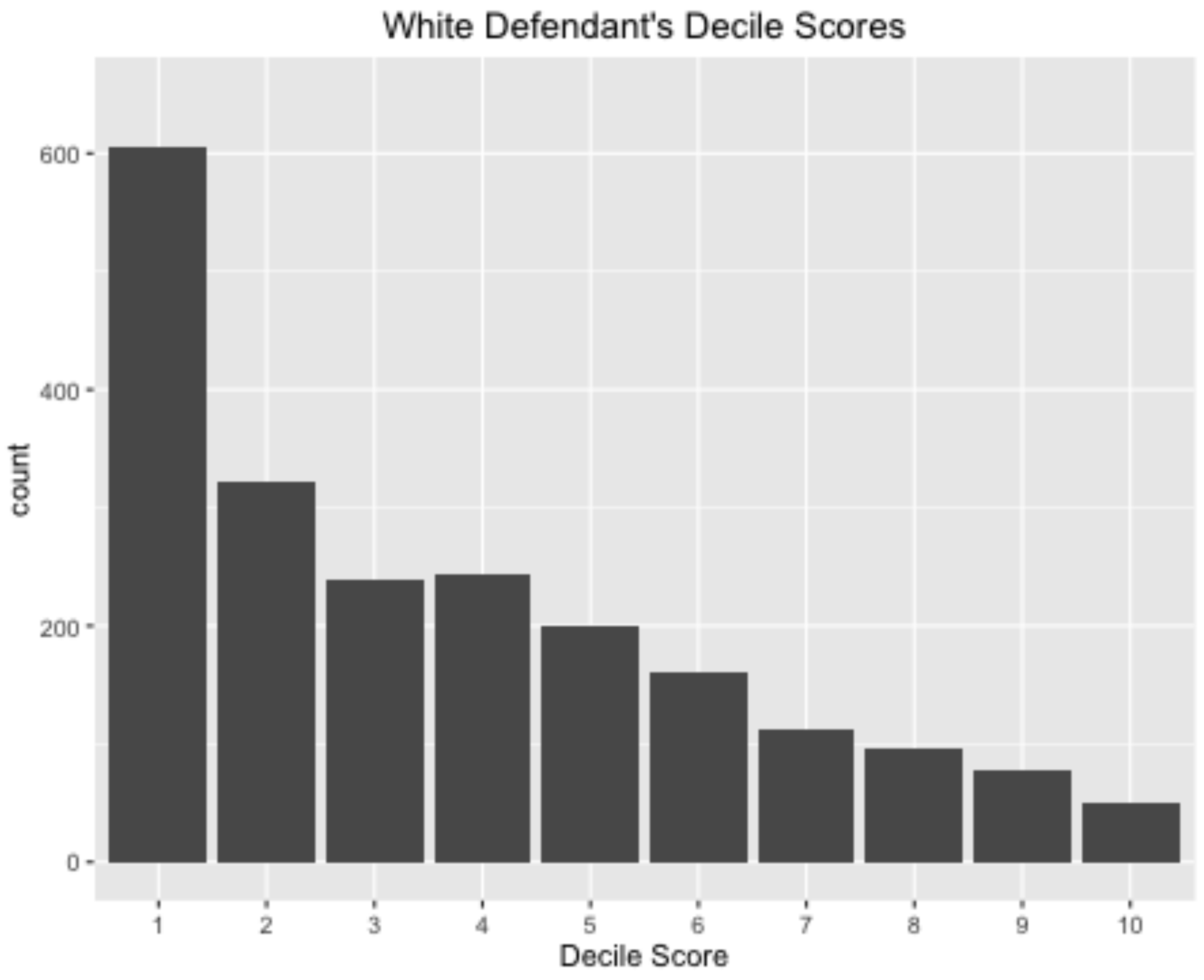
Engineering equitable software requires conscious effort

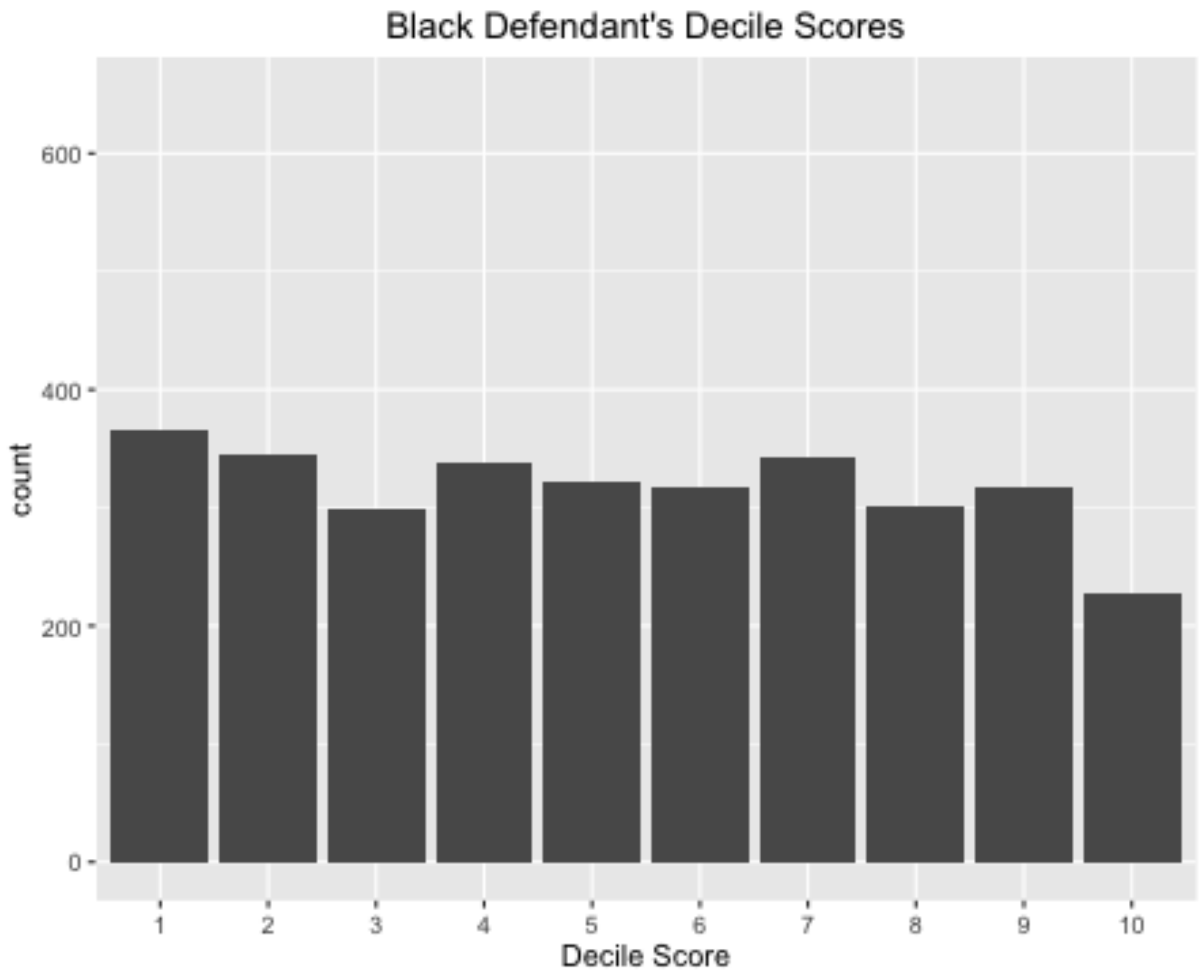
- How do we determine what “the right thing” is?
- How do we convince our investors/managers to take this action?
- Out of scope for this week: ethical frameworks (how do we reason between multiple compelling tradeoffs)
- This lesson: Discuss some examples that hopefully we all agree are problematic

Software Engineering & Ethics Example

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

Software Engineering & Ethics Example

Algorithmic Bias: Price Discrimination

THE WALL STREET JOURNAL.

Websites Vary Prices, Deals Based on Users' Information

SNAPSAFE; HOME DEPOT; ROSETTA STONE

By Jennifer Valentino-DeVries, Jeremy Singer-Vine and Ashkan Soltani

December 24, 2012

https://www.wsj.com/articles/SB10001424127887323777204578189391813881534

2017 IEEE European Symposium on Security and Privacy

FairTest: Discovering Unwarranted Associations in Data-Driven Applications*

Florian Tramèr[1], Vaggelis Atlidakis[2], Roxana Geambasu[2], Daniel Hsu[2], Jean-Pierre Hubaux[3], Mathias Humbert[4], Ari Juels[5], Huang Lin[3]

[1]*Stanford*, [2]*Columbia University*, [3]*EPFL*, [4]*Saarland University*, [5]*Cornell 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 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.

1. Introduction

Today's applications collect and mine vast quantities of personal information. Such data can boost applications' utility by personalizing content and recommendations, increase business revenue via targeted product placement, and improve a wide range of socially beneficial services, such as healthcare, disaster response, and crime prevention.

The collection and use of such data raise two important challenges. First, massive data collection is perceived by many as a major threat to traditional notions of individual privacy. Second, the use of personal data for algorithmic 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.

Staples' intentions aside, it is evidently difficult for programmers to foresee all the subtle implications and risks of data-driven heuristics. Moreover, these risks will only increase as data is passed through increasingly complex machine learning (ML) algorithms whose associations and inferences may be impossible to anticipate.

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 [5].

We refer to such bugs generically 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].

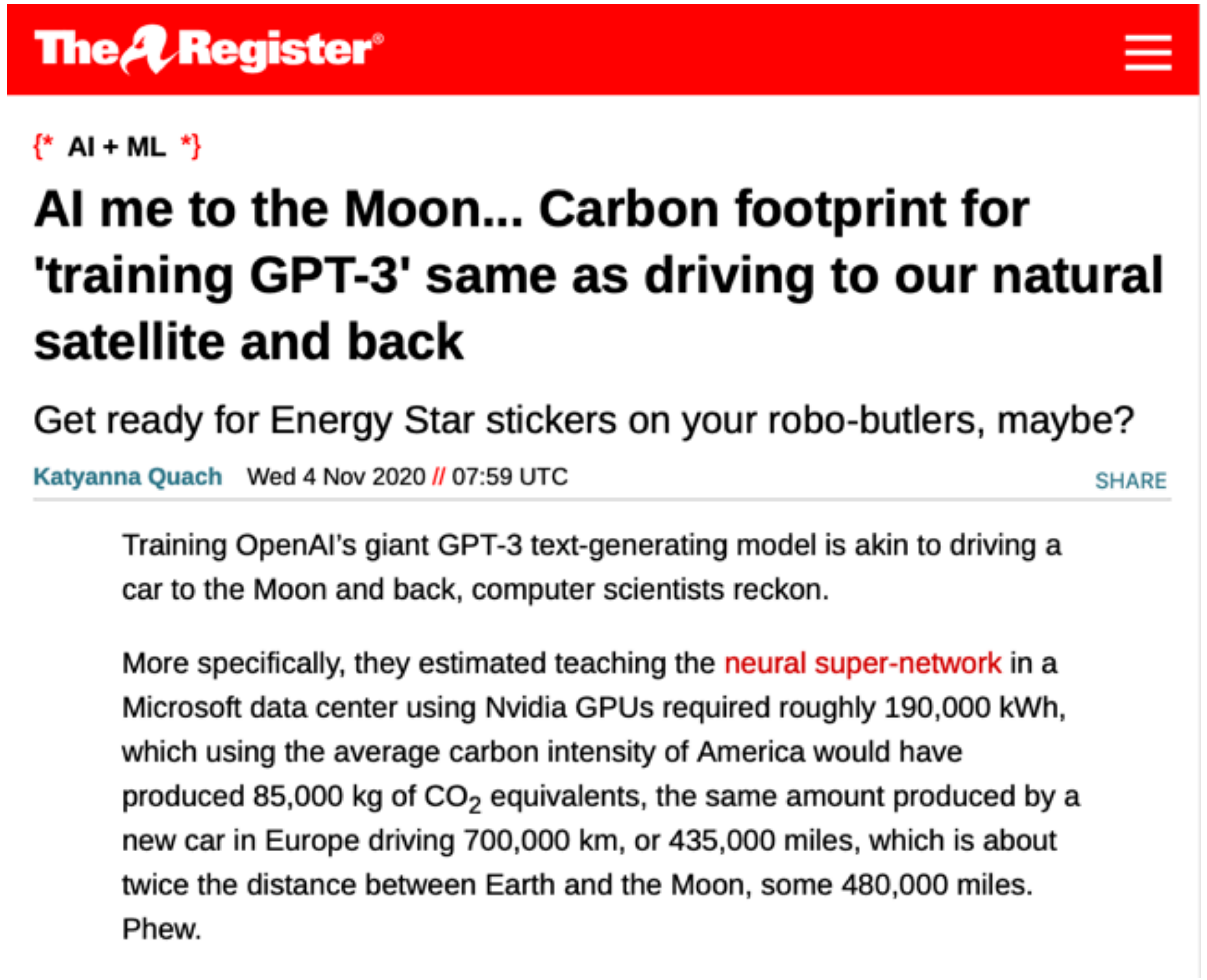
The Unwarranted Associations Framework. In order to

*Work done while the first author was at EPFL.

DOI 10.1109/EuroSP.2017.29

Software Engineering & Ethics Example

Climate Impact: Machine Learning Model Training & Development



The Register

{* 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 CO_2 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_2e (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" Emma Strubell, Ananya Ganesh, Andrew McCallum, in Proceedings of ACL 2019

Software Engineering & Ethics Example

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

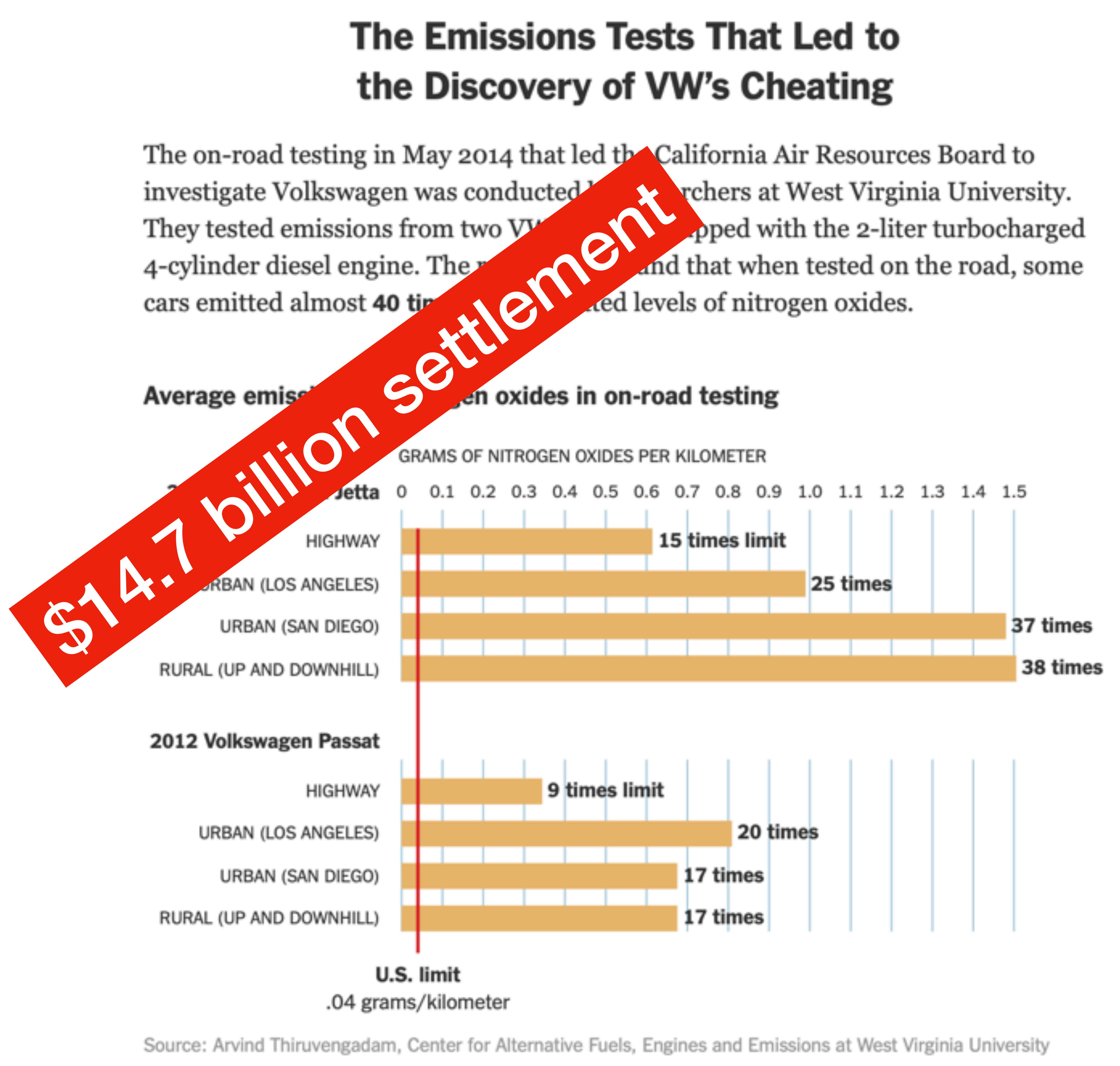
SHARE

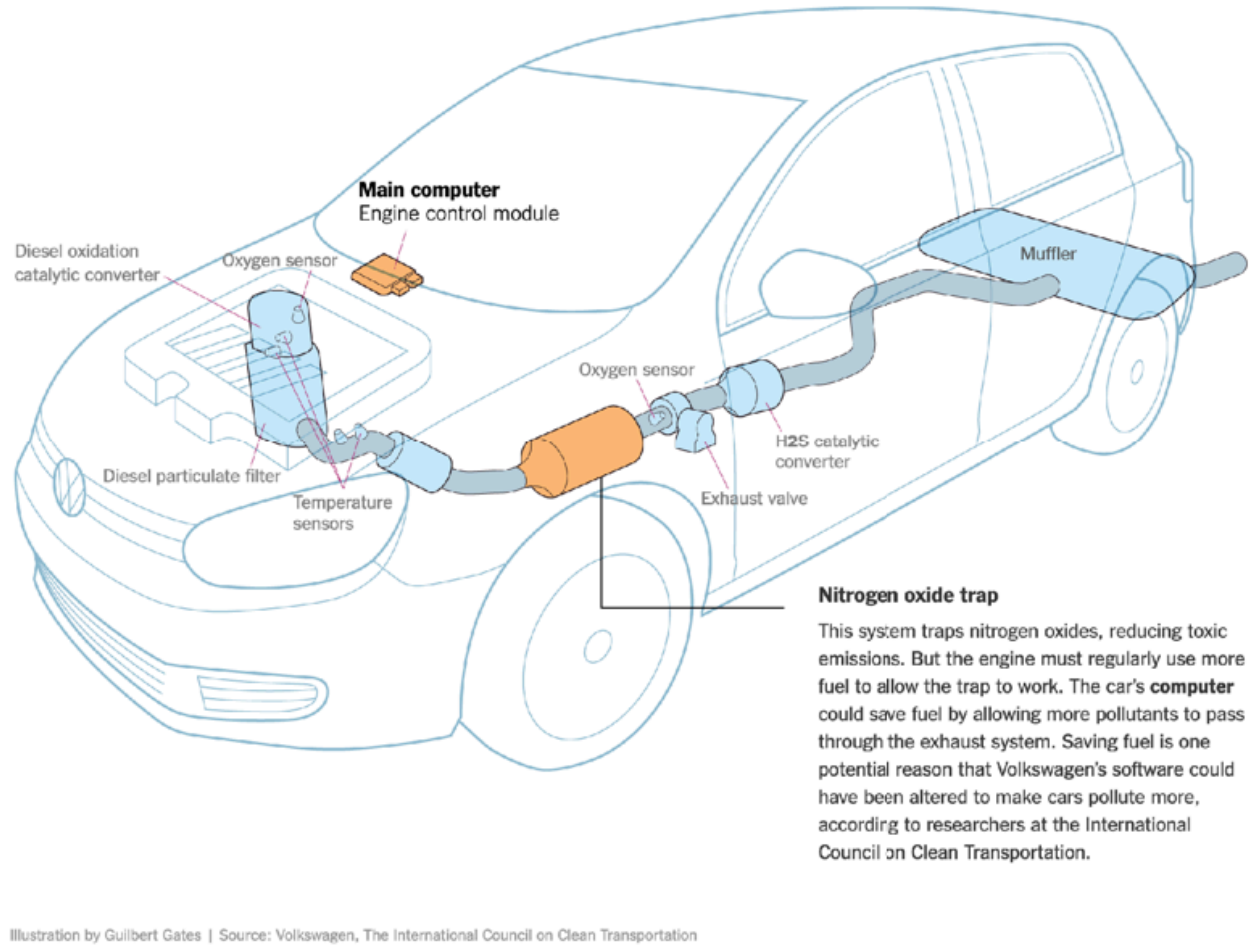
Jul 15 2019	**Brief amicus curiae of Washington Legal Foundation filed.**
Jul 15 2019	**Brief amici curiae of Retail Litigation Center, Inc., et al. filed.**
Jul 15 2019	**Brief amicus curiae of Cato Institute filed.**
Jul 15 2019	**Brief amicus curiae of Restaurant Law Center filed.**
Jul 15 2019	**Brief 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

Software Engineering & Ethics Example

Evading regulation: Volkswagen





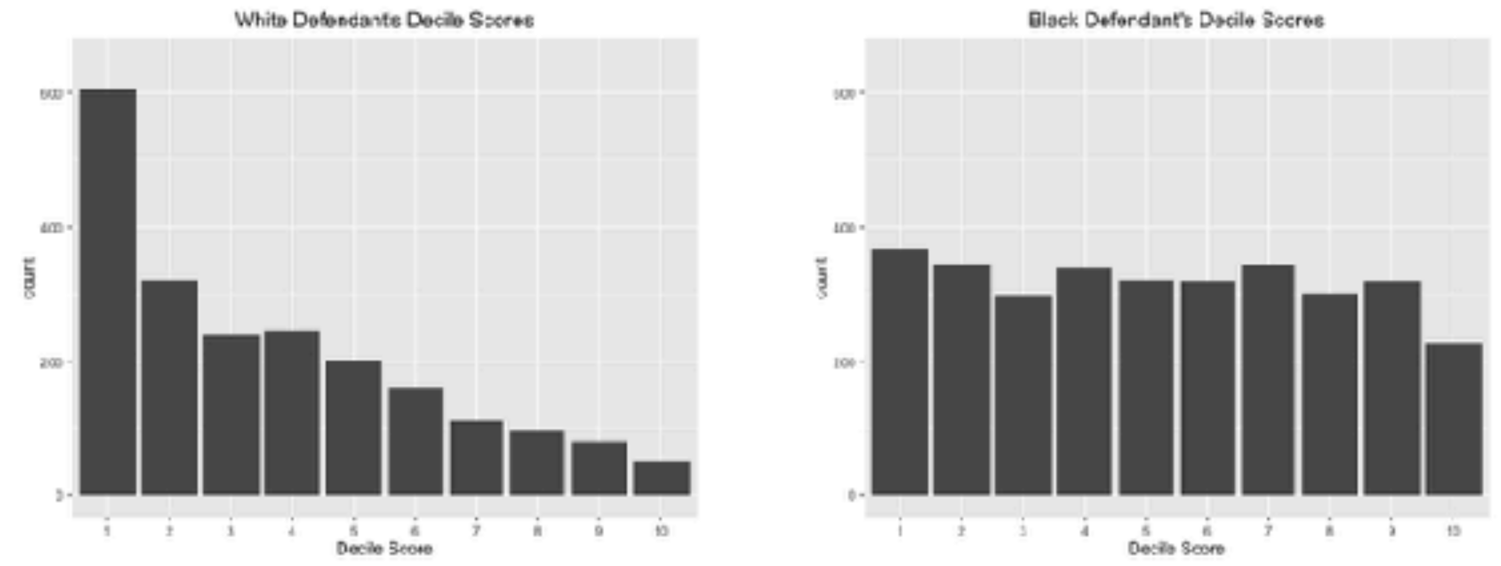
"How Volkswagen's 'Defeat Devices' Worked" By Guilbert Gates, Jack Ewing, Karl Russell and Derek Watkins

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

https://www.wsj.com/articles/SB10001424127887323777204578189391813881534

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

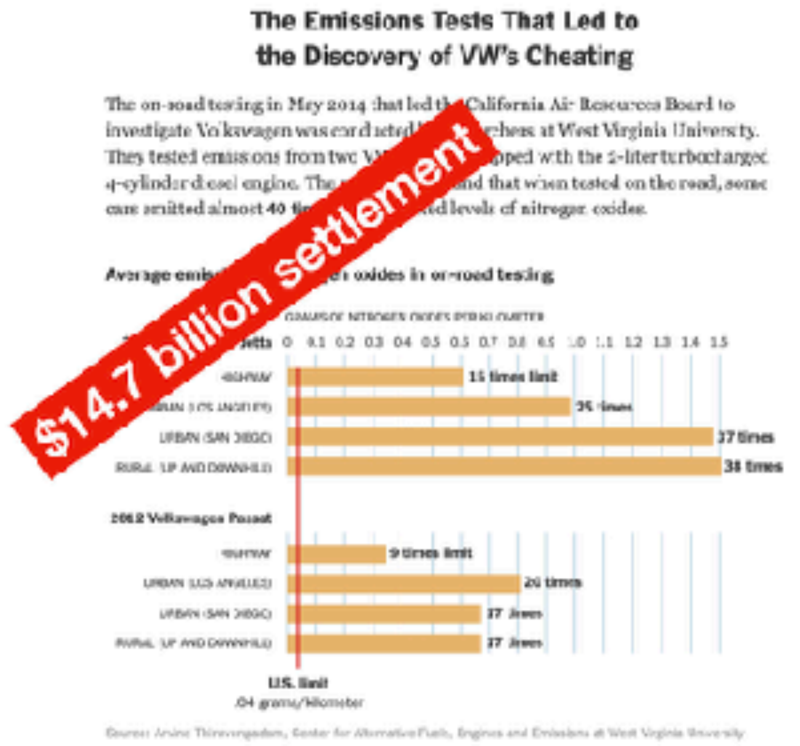
Domino's Would Rather Go to the Supreme Court Than Make Its Website Accessible to the Blind

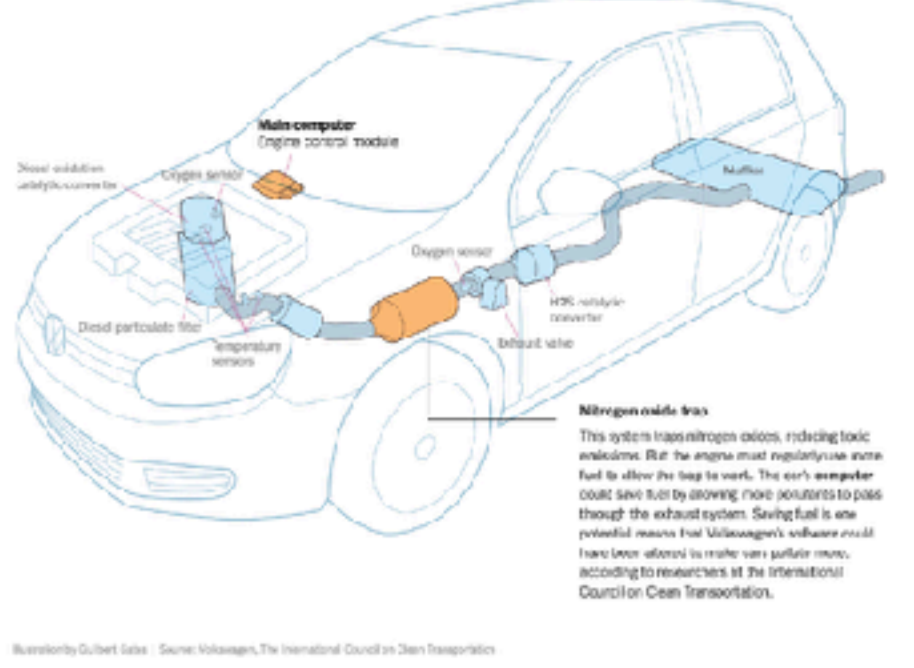
Rather than developing technology to support users with disabilities, the pizza chain is taking its fight to the top

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

Engineering Equitable Software

This week's roadmap

- This lesson: What does it mean to build software that is equitable?
- 11.2: Ethics in Software Engineering
- 11.3: Acceptance & Inclusivity Testing