CS 4530
Fundamentals of Software Engineering

Module 17: Engineering Software for Equity

Adeel Bhutta, Mitch Wand (with contributions by Christo Wilson)

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

Ethically and morally implicated technology is **everywhere**!

- Algorithms that gate access to loans, insurance, employment, government services…
- Algorithms that perpetuate or exacerbate existing discrimination
- Bad medical software can kill people (Therac-25)
- UIs that discriminate against differently-abled people (Domino's)
- Third-party data collection for hyper-targeted advertising
- GPT-3 !!
- And on… and on… and on…

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.

Quote: "Software Engineering at Google: Lessons Learned from Programming Over Time," Wright et al, 2020 (O'Reilly)

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

Quote: “Software Engineering at Google: Lessons Learned from Programming Over Time,” Wright et al, 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?

An approach: consider human values in the design process.

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

Engaging with values in the design process offers creative opportunities for:

- Technical innovation
- Improving the human condition (*doing good* and *saving the world*)

This is called *Value-Sensitive Design*

- An approach championed by some folks at the University of Washington
- The next few slides are adapted from the ones at vsd.ccs.neu.edu (Thank you, Christo Wilson and your collaborators).
- More resources linked on the module page.

Challenges: how to

- Define success objectives?
- Identify the social structure in which a technology is situated?
- Identify legitimate direct and indirect stakeholders?
- Elicit the full range of values at play?
- Balance and address tensions between different values?
- Identify and mitigate unintended consequences?

Defining Success

In CS, we typically think about technical success

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

Example metrics:

- 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?

Example metrics:

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

Identifying Stakeholders

Direct Stakeholders

The sponsor (your employer, etc.)

Members of the design team

Demographically diverse users

- Races and ethnicities, men and women, LGBTQIA, differently abled, US vs. non-US, …

Special populations

- Children, the elderly, victims of intimate partner violence, families living in poverty, the incarcerated, indigenous peoples, the homeless, religious minorities, non-technology users, celebrities

Roles

- Content creators, content consumers, power users, …

Indirect Stakeholders

Bystanders

- Those who are around your users
- E.g. pedestrians near an autonomous car

“Human data points”

- Those who are passively surveilled by your system

Civil society

- E.g. people who aren’t on social media are still impacted by disinformation
- People who care deeply about the issues or problem being addressed

Those without access

- Barriers include: cost, education, availability of necessary hardware and/or infrastructure, institutional censorship…

Whose values are impacted by a piece of technology?

Filtering Stakeholders

It is tempting to be overly comprehensive when enumerating stakeholders…

But not every impacted individual has legitimate values at play

Examples:

- Foreign election meddlers are affected by content moderation, want to protect their “free speech”
- Dictatorships are impacted by universal encryption, want unfettered surveillance capabilities
- Cyber criminals want to steal things, are against cybersecurity measures

These stakeholders are not legitimate, may be safely ignored

Identifying the Full Range of Values

- Some values are universal: accessibility, justice, human rights, privacy
- Others are tied to specific stakeholders and social contexts
- Identifying relevant values:
  - Start with a thorough understanding of the relevant features of the social situation
  - Add experience/knowledge from similar technologies or design decisions (case studies, etc.)
  - Add results of empirical investigation
- What are the scale of impacts to various stakeholders?

Example Values

Human welfare refers to people's **physical,** material, and psychological well-being

Accessibility refers to making all people successful users of information technology

Respect refers to treating people with politeness and consideration

Calmness refers to a peaceful and composed psychological state

Freedom from bias refers to systematic unfairness perpetrated on individuals or groups, including pre-existing social bias, technical bias, and emergent social bias

More Example Values

Ownership and property refers to a right to possess an object (or information), use it, manage it, derive income from it, and bequeath it

Privacy refers to a claim, an entitlement, or a right of an individual to determine what information about himself or herself can be communicated to others

Trust refers to expectations that exist between people who can experience good will, extend good will toward others, feel vulnerable, and experience betrayal

Accountability refers to the properties that ensures that the actions of a person, people, or institution may be traced uniquely to the person, people, or institution

Addressing Value Tensions

The most challenging step in VSD, by far

- This is where the hard choices happen

What are the core values that cannot be violated?

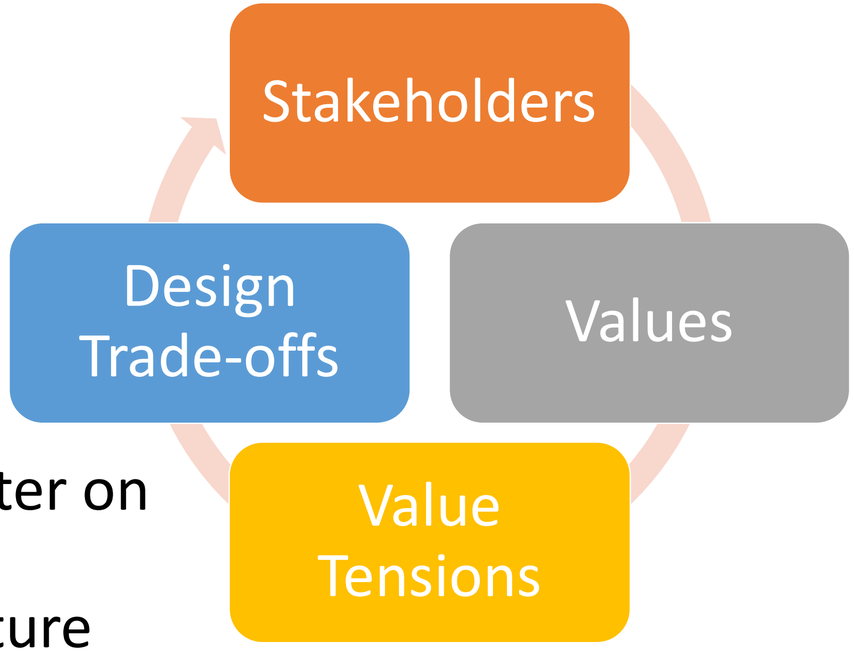
Which tensions can be addressed through:

- Technological mechanisms?
- Social mechanisms?

When a tension cannot be reconciled, whose values take precedence?

What tensions must be addressed immediately, versus later on through additional features?

- Early design decisions will unavoidably foreclose future design possibilities



Identifying Unintended Consequences

- Technology *will* be adopted in unanticipated ways. Being intellectually rigorous means considering and mitigating risks in designs ahead of time.
- What if:
  - Our recommendation system promotes misinformation or hate speech?
  - Our database is breached and publicly released?
  - Our facial recognition AI is used to identify and harass peaceful protestors?
  - Our child safety app is used to stalk women?
  - Our chatbot is sexist or racist?

Example: Content Moderation

The issue: *free expression* in tension with *welfare* and *respect*

- Some speech may be hurtful and/or violent
- Removing this speech may be characterized as censorship

Bad take: unyielding commitment to free speech, no moderation

- Trolls and extremists overrun the service, it becomes toxic, all other users leave
- Violent speech actually impedes free speech in general

Bad take: strict whitelists of acceptable speech

- Precludes heated debate, discussion of “sensitive topics”
- Disproportionately impacts already marginalized groups

Good take: recognizing that moderation will never be perfect, there will be mistakes and grey areas

- Doing nothing is not a viable option
- Clear guidelines that are earnestly enforced create a culture of accountability

Strategies for Addressing Value Tensions

Identify Red Lines	*Red lines*: bedrock values that cannot be violated • Address these first
Look for Win—Wins	Look for win-win scenarios • Some stakeholders may be agreement; others may want the same outcome but for different reasons
Embrace Tradeoffs	Be open and honest when value tradeoffs are necessary • E.g. when functionality and privacy are in tension, both can be addressed through informed consent
Don’t Forget Social Solutions	Creatively leverage technical and social solutions in concert • E.g. if a new system is going to automate away jobs, pair it with a retraining program

Practical Tips

Adopt, Extend, Adapt	Adopt and extend the methods for your own purposes. Adapt for your sociotechnical setting.
Variety	Use a variety of empirical values-elicitation methods, rather than relying on a single one.
Continuous Evaluation	Continue to elicit stakeholder values throughout the design. If new values of import surface during the design process, engage them.
Anticipation	Anticipate unanticipated consequences: continue the VSD process throughout the deployment of the technology
Collaborate	Particularly with people from other disciplines, and those with deep contextual knowledge of and expertise in your sociotechnical setting.

Where does this leave us?

- **So that we can sleep at night**
  - Consider the different ways that our software may impact others
  - Consider the ways in which our software interacts with the political, social, and economic systems in which we and our users live
  - Follow best practices, and actively push to improve them
  - Encourage diversity in our development teams
  - Engage in honest conversations with our co-workers and supervisors to explore possible ethical issues and their implications.

After this point are the old slides

Code of Ethics

Professional Engineers

Engineers, in the fulfillment of their professional duties, shall:

1. Hold paramount the safety, health, and welfare of the public.
2. Perform services only in areas of their competence.
3. Issue public statements only in an objective and truthful manner.
4. Act for each employer or client as faithful agents or trustees.
5. Avoid deceptive acts.
6. Conduct themselves honorably, responsibly, ethically, and lawfully so as to enhance the honor, reputation, and usefulness of the profession.

National Society of Professional Engineers - Code of Ethics for Engineers

Code of Ethics

Professional Engineers: Citigroup Center

- Design met building code, but did *not* account for all failure modes
- Last-minute changes to construction increased odds of failure
- Fixed before disaster could strike, but kept a secret for 20 years

https://en.wikipedia.org/wiki/Citicorp_Center_engineering_crisis

Badly-engineered software can kill people

Therac-25 (1985-1987)

- Bug in software caused 100x greater exposure to radiation than intended
- At least 6 died
- Likely far more suffered: deaths occurred over a period of 2 years!
- Weak accountability in manufacturer's organization

"Therac-25" by Catalina Márquez, Wikimedia commons, CC BY-SA 4.0

https://ethicsunwrapped.utexas.edu/case-study/therac-25

ACM’s Code of Ethics Software Engineers

1. PUBLIC – Software engineers shall act consistently with the public interest.

2. CLIENT AND EMPLOYER – Software engineers shall act in a manner that is in the best interests of their client and employer consistent with the public interest.

1. PUBLIC – Software engineers shall act consistently with the public interest.

5. MANAGEMENT – Software engineering managers and leaders shall subscribe to and promote an ethical approach to the management of software development and maintenance.

6. PROFESSION – Software engineers shall advance the integrity and reputation of the profession consistent with the public interest.

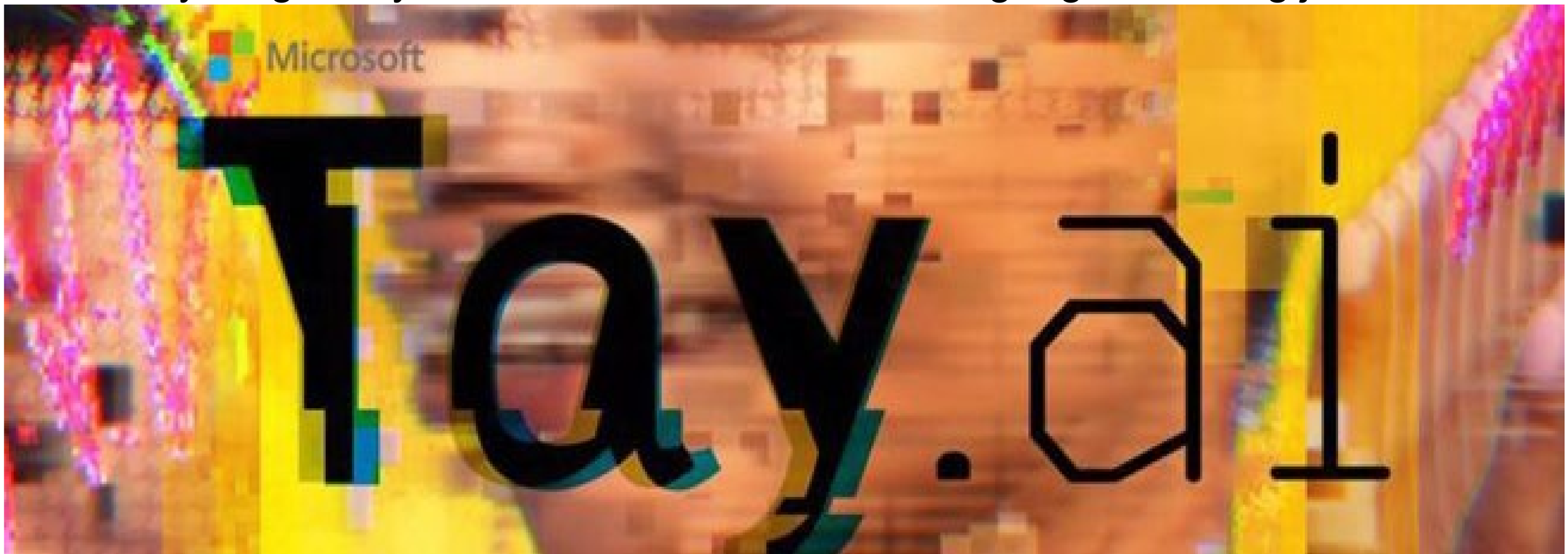
7. COLLEAGUES – Software engineers shall be fair to and supportive of their colleagues.

8. SELF – Software engineers shall participate in lifelong learning regarding the practice of their profession and shall promote an ethical approach to the practice of the profession.

https://ethics.acm.org/code-of-ethics/software-engineering-code/

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%**

Analysis of Broward County data: "How We Analyzed the COMPAS Recidivism Algorithm" by Larson et al.

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

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

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

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 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

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" by Strubell et al in ACL19

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 of America, et al. filed.**

Software evades regulation

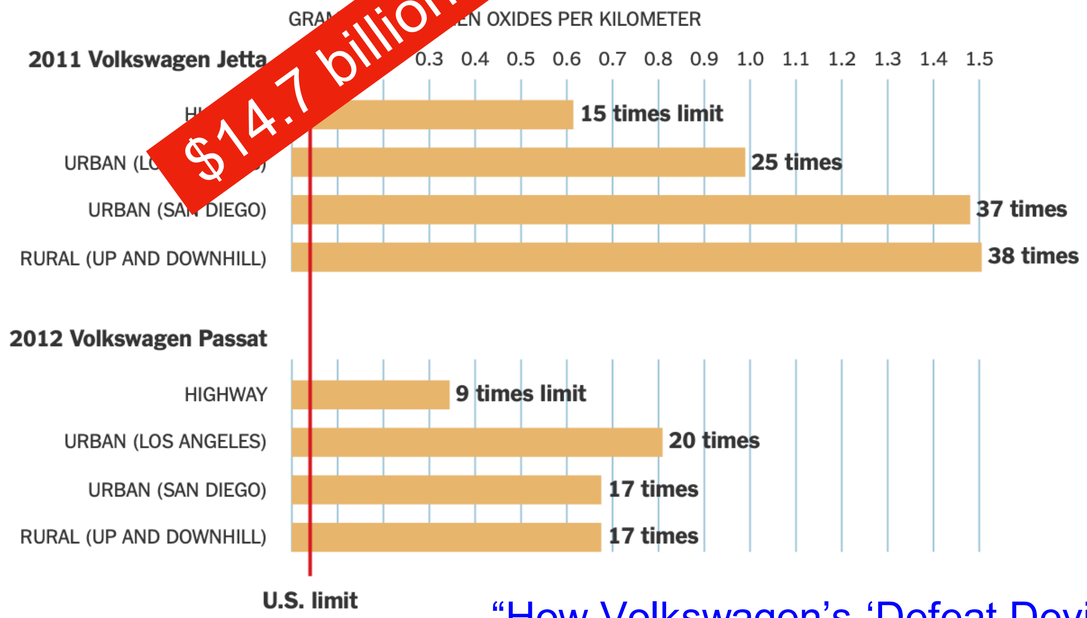
Example: Volkswagen diesel emissions

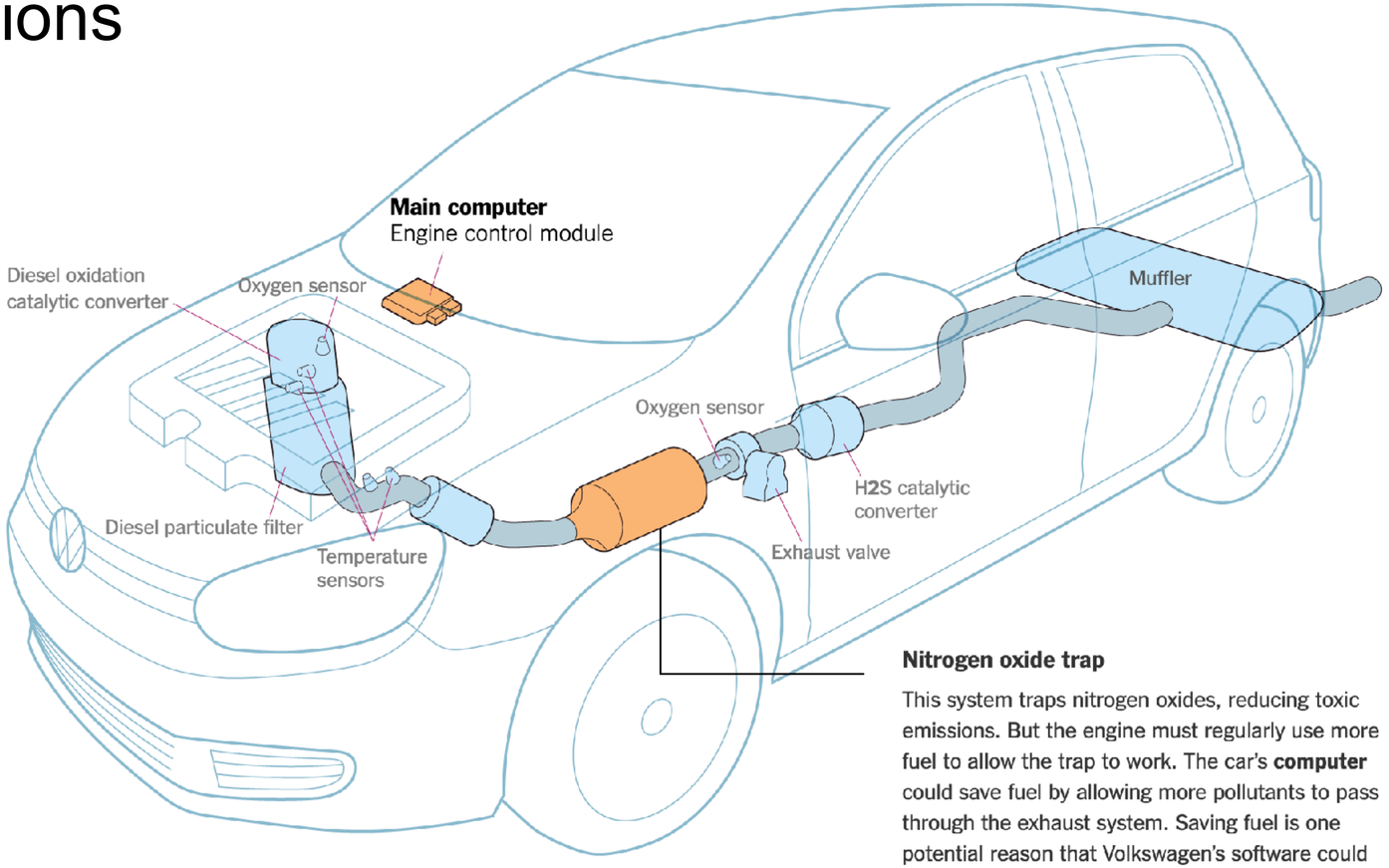
The Emissions Tests That Led to the Discovery of VW's Cheating

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 th… …en tested on the road, some cars emitted almost **40 times** the permitted l… …rogen oxides.

Average emissions of nitrogen ox… …-road testing

$14.7 billion settlement





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

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

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

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.

In WIRED's tests, Google Photos did identify some primates, but no gorillas like this one were to be found. RICK MADONIK/TORONTO STAR/GETTY IMAGES

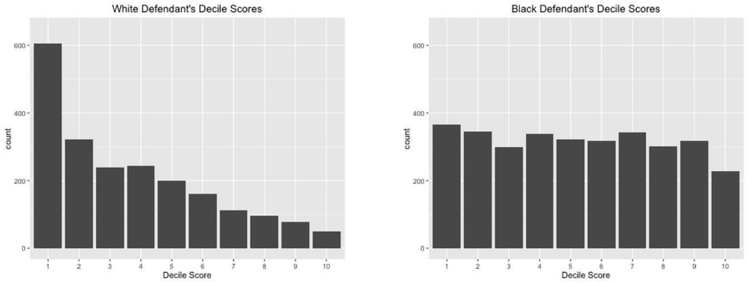
https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Reflecting on these examples

Personal philosophies and business cases

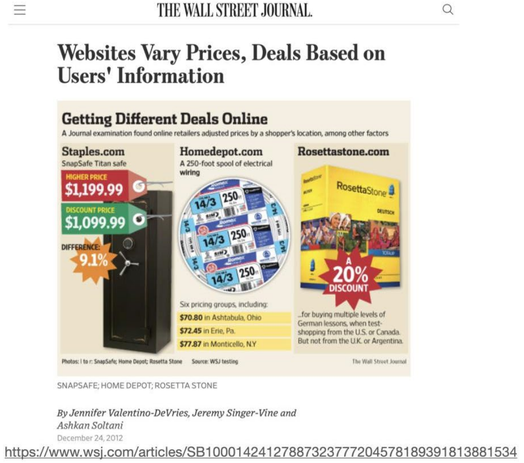
Algorithmic Bias: COMPAS Sentencing Tool

	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
...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: (to r: SnapSafe; Home Depot; Rosetta Stone Source: WSJ testing The Wall Street Journal

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

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

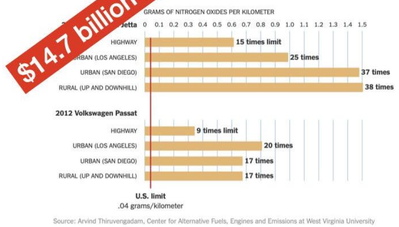
Evading regulation: Volkswagen

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 VWs equipped with the 2-liter turbocharged 4-cylinder diesel engine. They found that when tested on the road, some cars emitted almost 40 times the allowed levels of nitrogen oxides.

Average emissions of nitrogen oxides in on-road testing

$14.7 billion settlement



Source: Arvind Thiruvengadam, Center for Alternative Fuels, Engines and Emissions at West Virginia University

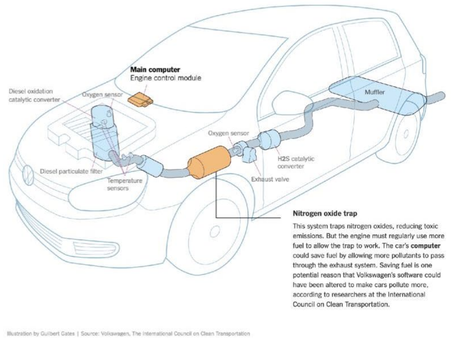


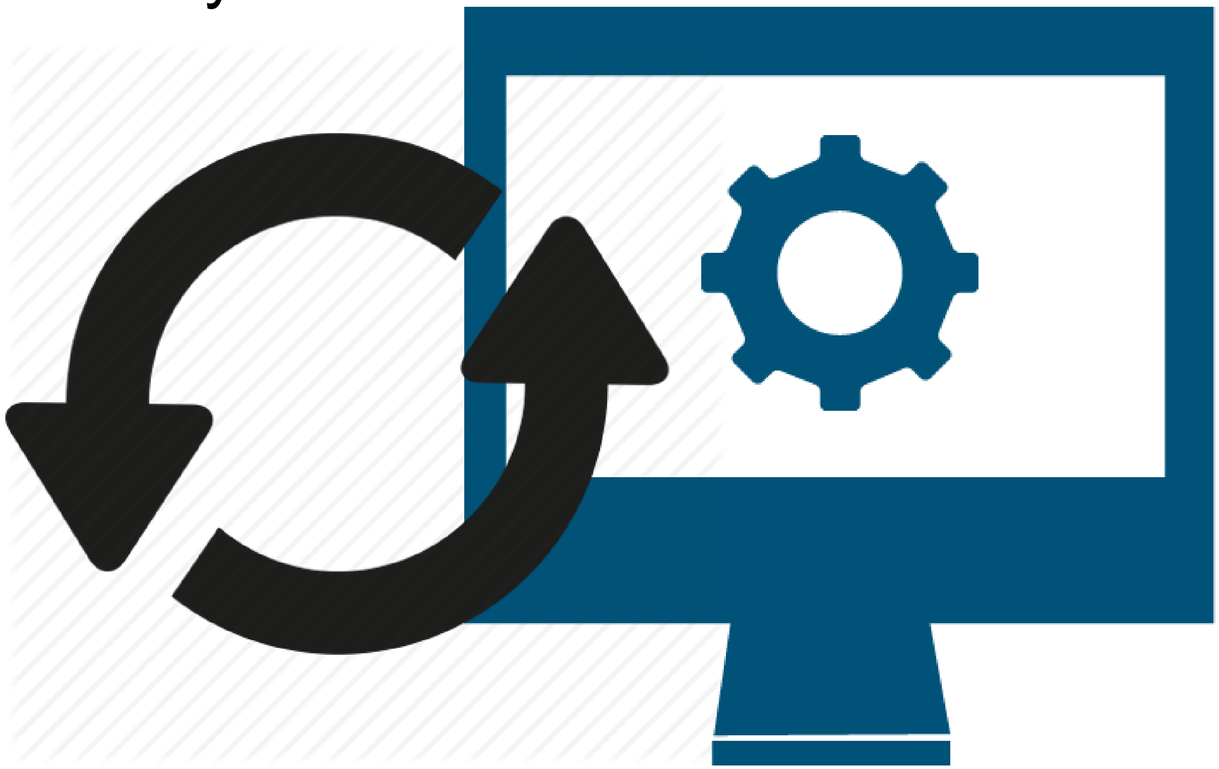
Illustration by Guilbert Gates | Source: Volkswagen, The International Council on Clean Transportation

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

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

https://www.microsoft.com/en-us/research/uploads/prod/2020/03/Guidelines_summary_image@2x.png

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