

17-423/723: Designing Large-scale Software Systems

Ethical & Responsible Design

April 16, 2025

Logistics

- Project presentations next Wednesday (last lecture)
 - Presentation instructions out today
 - **Everyone must be present**
- No recitation this Friday
 - Use this time to work on the presentation!

Learning Goals

- Describe ethical responsibilities of a software engineer
- Identify different types of harms on the society that can be caused by a software system
- Consider and select an appropriate type of fairness for a system being designed
- Consider whether the benefits of building a software product outweighs its potential harm

Responsible Software Engineering



In 2015, Shkreli received widespread criticism [...] obtained the manufacturing license for the antiparasitic drug Daraprim and raised its price from USD 13.5 to 750 per pill [...] referred to by the media as "the most hated man in America" and "Pharma Bro". - Wikipedia

"I could have raised it higher and made more profits for our shareholders. Which is my primary duty." - Martin Shkreli

Terminology

- **Legal:** In accordance to societal laws
 - Systematic body of rules governing society; set by the government
 - Punishment for violation
- **Ethical:** Following moral principles of tradition, group, or individual
 - Branch of philosophy, science of a standard human conduct
 - Professional ethics: Rules codified by professional organization
 - No legal binding; no enforcement beyond "shame"
 - High ethical standards may yield long term benefits through image and staff loyalty

With a few lines of code...

- Software engineers have significant power in shaping products
- Even small design decisions can have substantial impact on users, the environment, and society (safety, security, privacy, discrimination...), even if not always intended
- **Our viewpoint:** We have both legal & ethical responsibilities to anticipate possible misuses of a system that we build, think through their consequences, and design mitigations against them

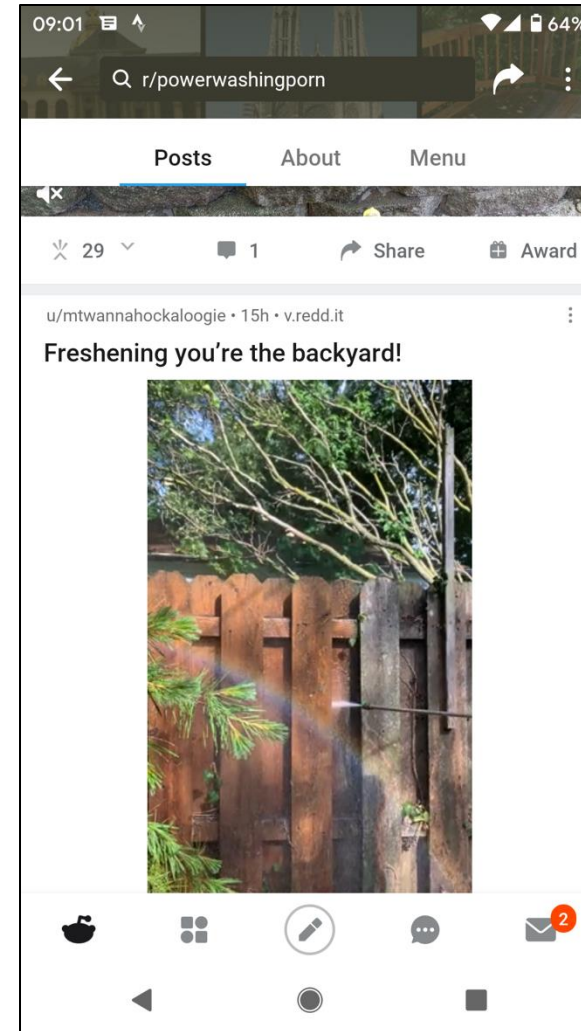
Example: Social Media



- **Q. What is the (real) objective of the organization?**

Optimizing for Organizational Objective

- How do we maximize the user engagement?
- **Examples:**
 - Infinite scroll: Encourage non-stop, continual use
 - Personal recommendations: Suggest news feed to increase engagement
 - Push notifications: Notify disengaged users to return to the app
- **Q. So what? How can social media harm people?**



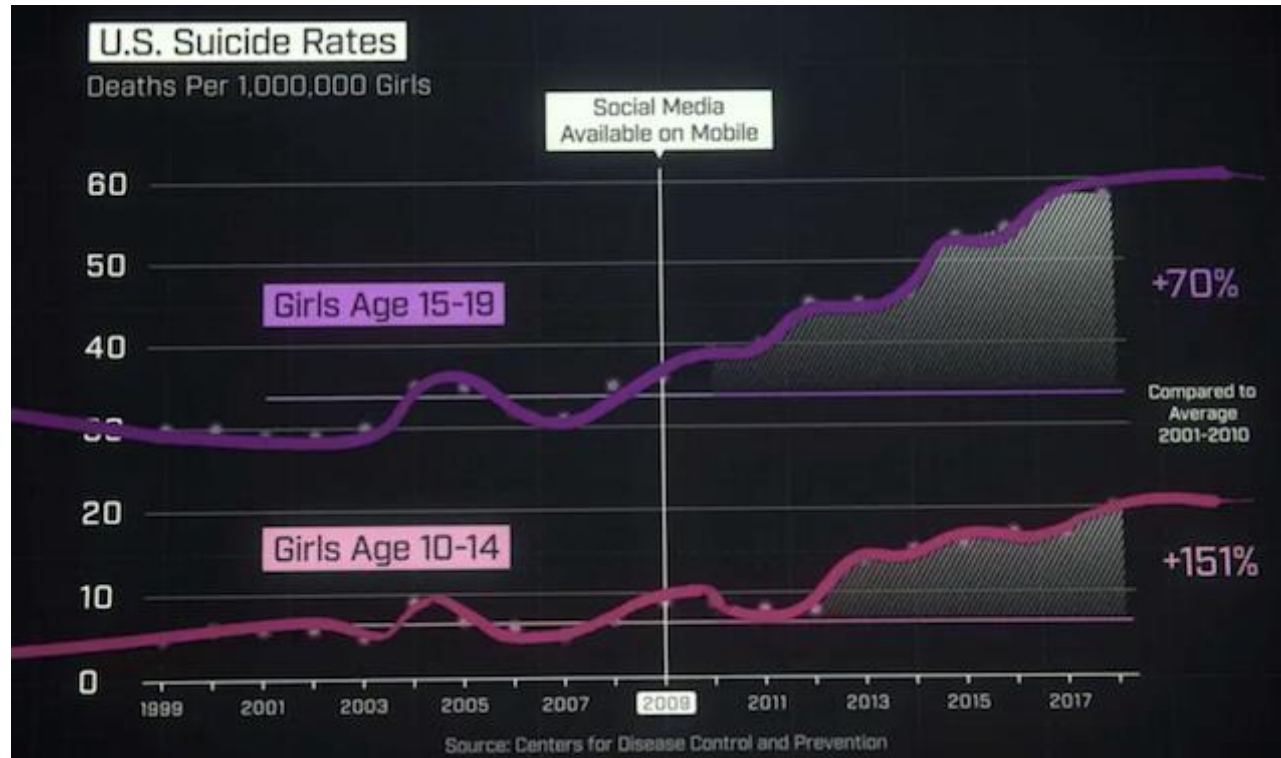
Addiction



- 210M people worldwide addicted to social media
- 71% of Americans sleep next to a mobile device
- ~1000 people injured per day due to distracted driving (US)

Sources: <https://www.flurry.com/blog/mobile-addicts-multiply-across-the-globe>
https://www.cdc.gov/motorvehiclesafety/Distracted_Driving/index.html

Mental Health



- 35% of US teenagers with low social-emotional well-being have been bullied on social media.
- 70% of teens feel excluded when using social media.

Sources: <https://leftronic.com/social-media-addiction-statistics>

Disinformation & Polarization



Who is to blame?

GOOGLE QUIETLY REMOVES 'DON'T BE EVIL' PREFACE FROM CODE OF CONDUCT

Google employees resigned this month over the company's autonomous weapons project

Anthony Cuthbertson | @ADCuthbertson | Monday 21 May 2018 12:21



**Q. Are these companies intentionally trying to cause harm?
If not, what are the root cause of the problem?**

Challenges

- Misalignment between organizational goals & societal values
 - Financial incentives often dominate other goals ("grow or die")
- Hardly any regulation
 - Little legal consequences for causing negative impact (with some exceptions)
 - Poor understanding of socio-technical systems by policy makers
- Engineering challenges
 - Difficult to clearly define or measure ethical values
 - Difficult to anticipate all possible usage contexts
 - Difficult to prevent malicious actors from abusing the system

Liability?

THE SOFTWARE IS PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Responsible Software Engineering Matters!

- We, as software engineers, have substantial power in shaping products and outcomes
- Serious individual and societal harms are possible from (a) negligence and (b) malicious designs
 - Safety failures, mental health problems, weapon proliferation
 - Security & privacy violations
 - Manipulation, addiction, surveillance, polarization
 - Job loss, deskilling
 - Discrimination

Value-Sensitive Design

- As software engineers, our focus is typically on the technical success of a product
- But there are also ethical **values** that we should explicitly consider as part of the design



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

Dimensions of Responsible Design

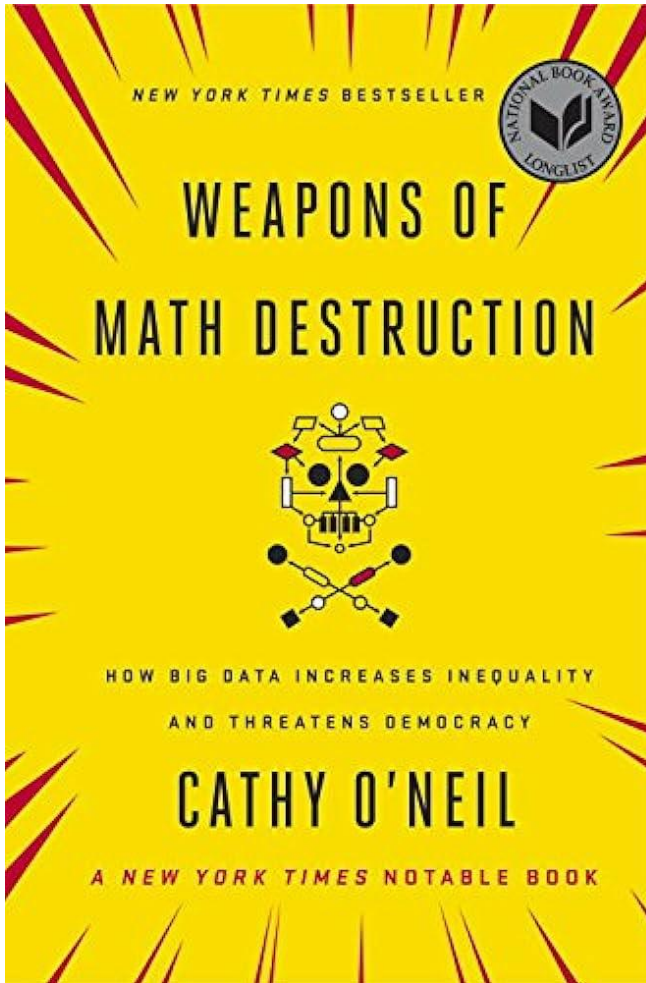
- Many quality attributes, when neglected, could result in harm to the stakeholders & environment
 - Robustness, reliability, safety
 - Security, privacy
 - Usability
 - Transparency, accountability
 - Accessibility
 - Fairness & bias - **today's focus!**
- We should deliberately consider these qualities and values during the design process

Fairness as Software Quality

What is fair?

Fairness discourse asks questions about how to treat people and whether treating different groups of people differently is ethical. If two groups of people are systematically treated differently, this is often considered unfair.

Why is fairness relevant for software?



- Increasing use of algorithms for making societal decisions
- Amplifying existing bias & discrimination against certain groups of population
- Many domains: College admission, job hiring, recidivism, loan lending, social networks...
- These problems have existed for a long time, but are being made worse by the rapid spread of ML

Types of Harms from Unfairness

- **Harms of allocation:** Withhold opportunities or resources from certain groups of population
- **Harms of representation:** Reinforce stereotypes and subordination along the lines of identity

Harms of Allocation

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

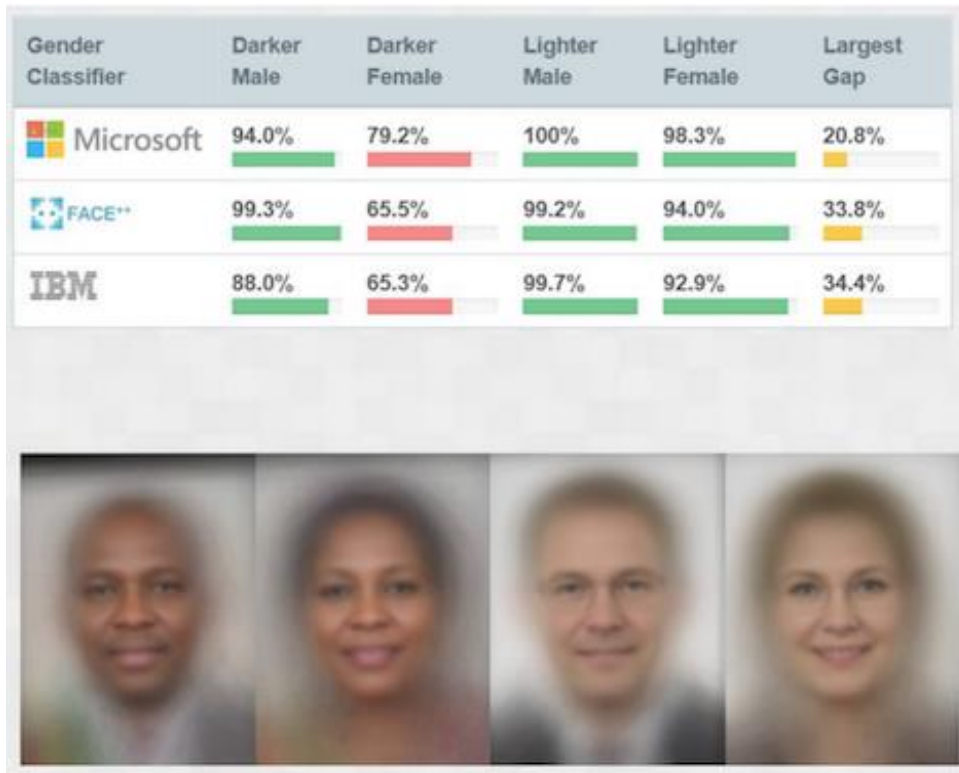
8 MIN READ



SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

- Withhold opportunities or resources from certain groups of people

Harms of Allocation



- Withhold opportunities or resources from certain groups of people
- Poor quality of service, degraded user experience for certain groups

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Buolamwini & Gebru (2018)

Harms of Representation

Ads by Google

[Latanya Sweeney, Arrested?](#)

1) Enter Name and State. 2) Access Full Background Checks Instantly.

www.instantcheckmate.com/

[Latanya Sweeney](#)

Public Records Found For: Latanya Sweeney. View Now.

www.publicrecords.com/

[La Tanya](#)

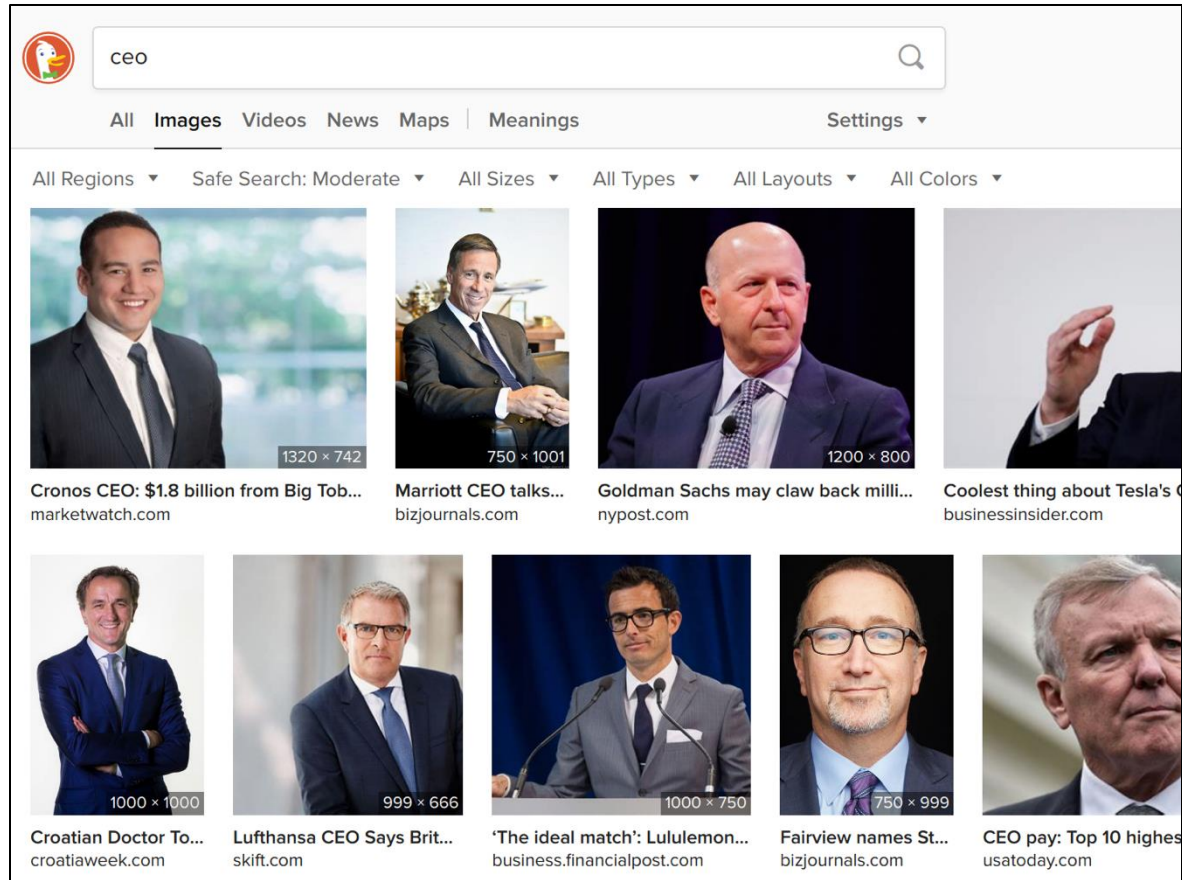
Search for La Tanya Look Up Fast Results now!

www.ask.com/La+Tanya

- Denigration: Unfair criticism of certain groups of individuals
- Reinforcement of existing stereotypes

Discrimination in Online Ad Delivery. Latanya Sweeney (2013)

Harms of Representation



- Over/under-representation of certain groups in organizations

Identifying Potential Harms

	Allocation of resources	Quality of Service	Stereotyping	Denigration	Over- / Under-Representation
Hiring system does not rank women as highly as men for technical jobs	x	x	x		x
Photo management program labels image of black people as “gorillas”		x		x	
Image searches for “CEO” yield only photos of white men on first page			x		x

- Multiple types of harms can be caused by a product!
- Think through and identify possible harms that can be caused

Challenges of incorporating algorithmic fairness into practice. Microsoft Research (2019)

Recall: What is fair?

Fairness discourse asks questions about how to treat people and whether treating different groups of people differently is ethical. If two groups of people are systematically treated differently, this is often considered unfair.

Discussion: “Fairly” dividing a pie

Which way is fair? Assume: Not everybody contributed equally during baking, and not everybody is equally hungry

- Equal slices for everybody
- Bigger slices for active bakers
- Bigger slices for inexperienced/new members (e.g., children)
- Bigger slices for hungry people
- More pie for everybody; bake more



Equality



The assumption is that **everyone benefits from the same supports**. This is equal treatment.

Equity



Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

Justice



All 3 can see the game without supports or accommodations because **the cause(s) of the inequity was addressed**. The systemic barrier has been removed.

Equality, Equity, or Justice?

- Equality
 - Treat everybody equally, regardless of their starting position
 - Focus on meritocracy; strive for fair opportunities
- Equity
 - Compensate for different starting positions
 - Lift disadvantaged groups (e.g., affirmative actions)
- Justice
 - Aspirational option; avoids a choice between equality and equity
 - Fundamentally removes initial imbalance or need for allocation decision
 - Involves rethinking the entire societal system in which the imbalance exists (in general, very challenging!)

Equality, Equity, or Justice?

Which way is fair? Assume: Not everybody contributed equally during baking, and not everybody is equally hungry

- Equal slices for everybody
- Bigger slices for active bakers
- Bigger slices for inexperienced/new members (e.g., children)
- Bigger slices for hungry people
- More pie for everybody; bake more



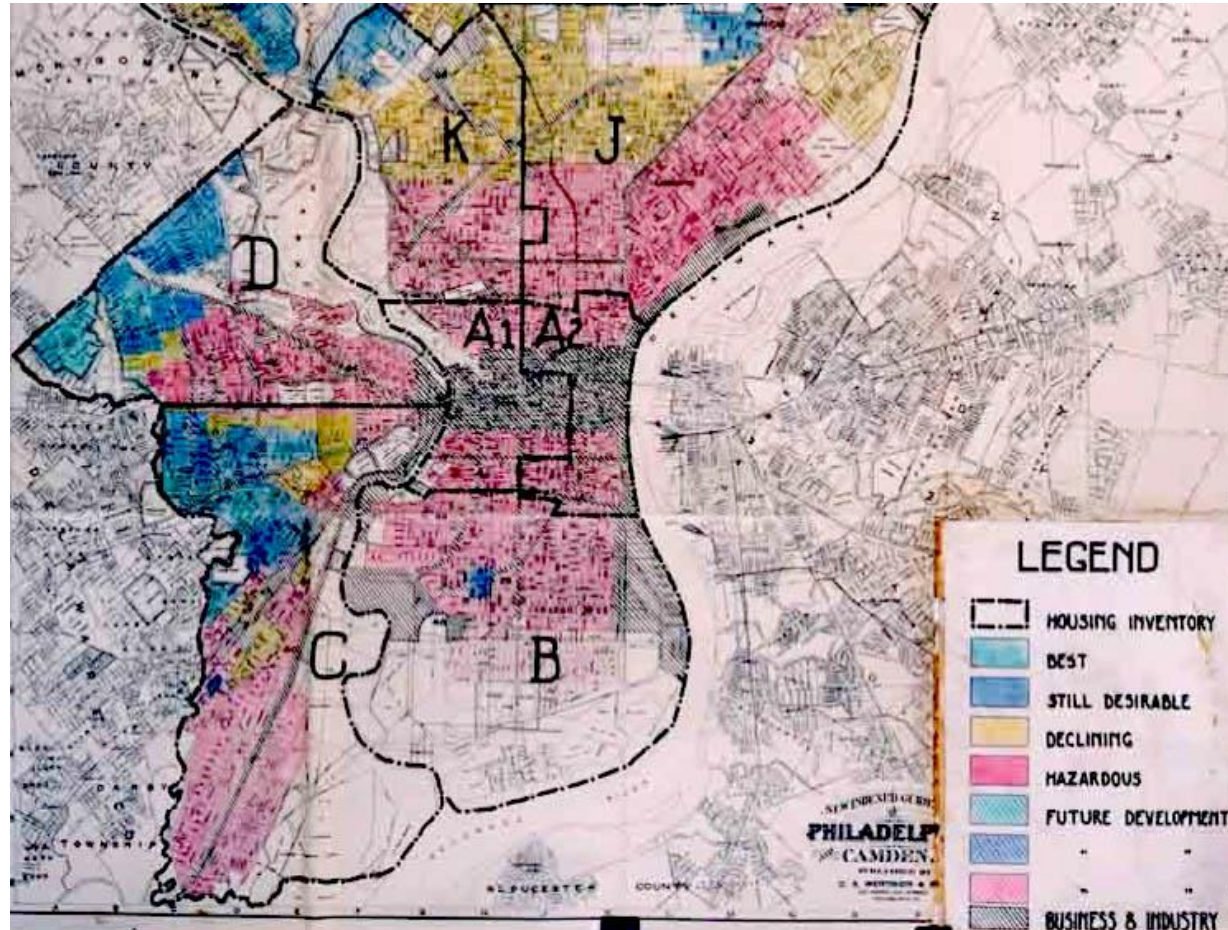
Discussion: Loan Lending Applications

Imagine developing an AI-based system for rating loan lending applications. What is fair?

- Distribute loans equally across population demographics
- Prioritize those who are more likely to pay back the loan (higher income, good credit history)
- Give special consideration to those from under-privileged backgrounds
- Give out loans to everyone



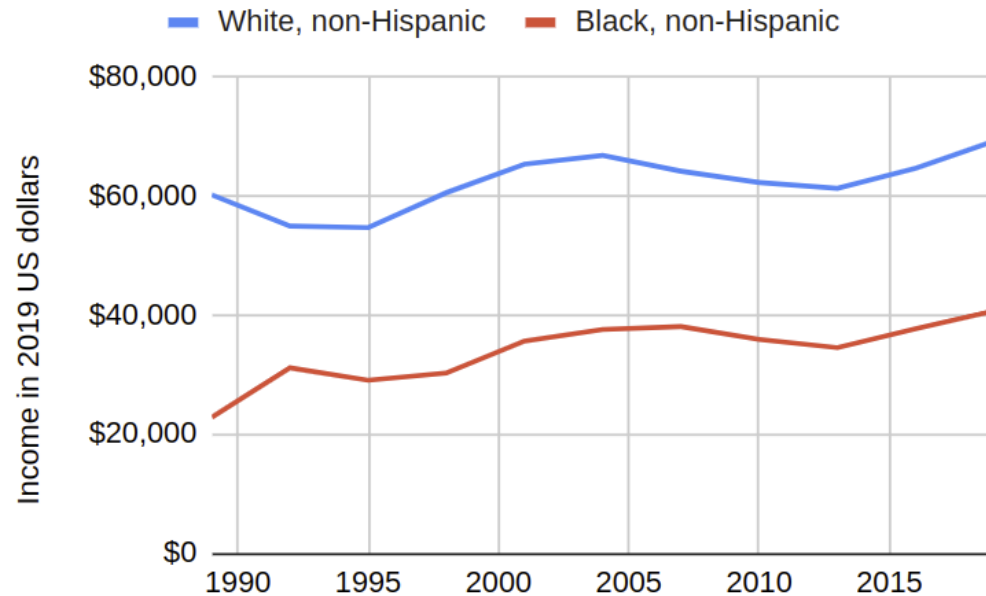
Redlining



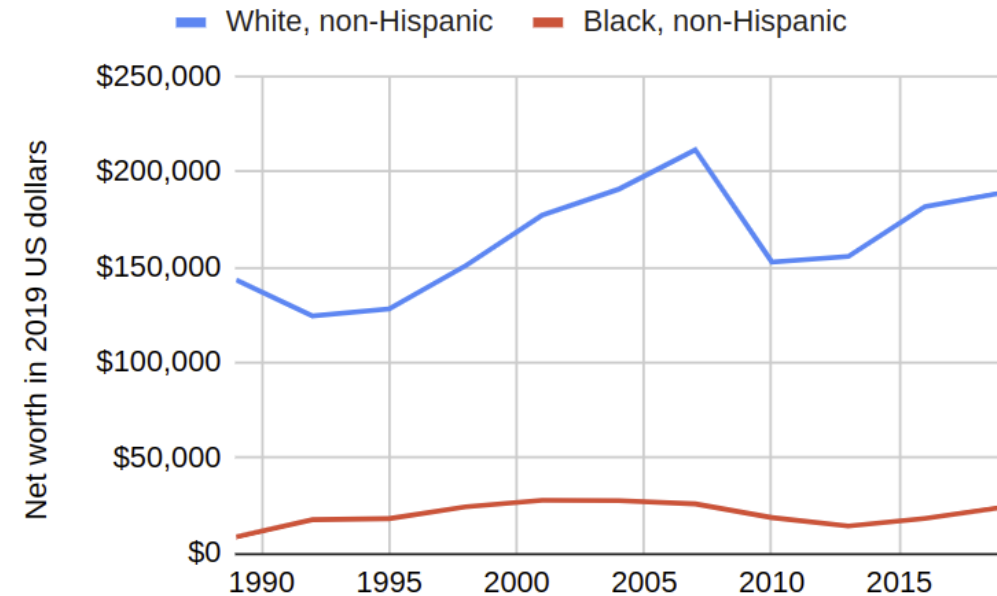
- Deliberately withhold services (e.g., mortgage, education, retail) from people in neighborhoods that are considered "risky"
- Map of Philadelphia, 1936, Home Owners' Loan Corps. (HOLC)
 - Classification based on estimated "riskiness" of loans
- Illegal practice; Equal Credit Opportunity Act (1974)

Historical bias, different starting positions

Median before-tax family income



Median family net-worth



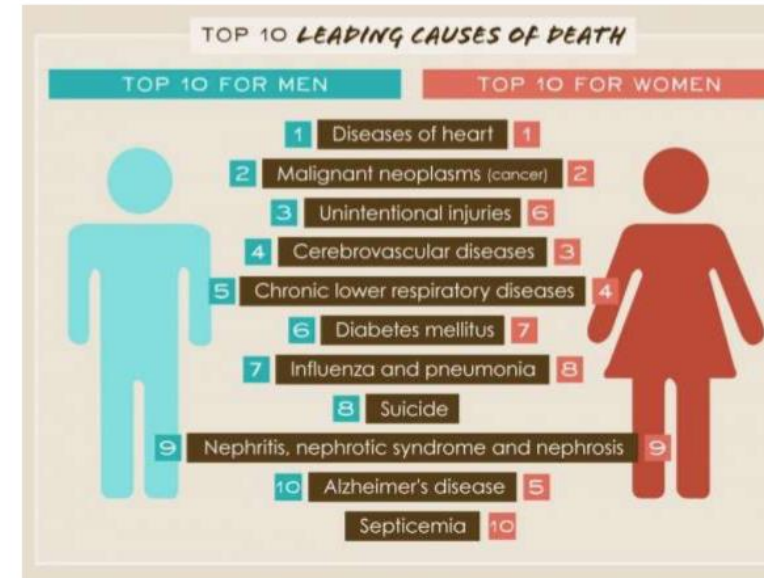
Source: [Federal Reserve's Survey of Consumer Finances](#)

Not all discrimination is harmful



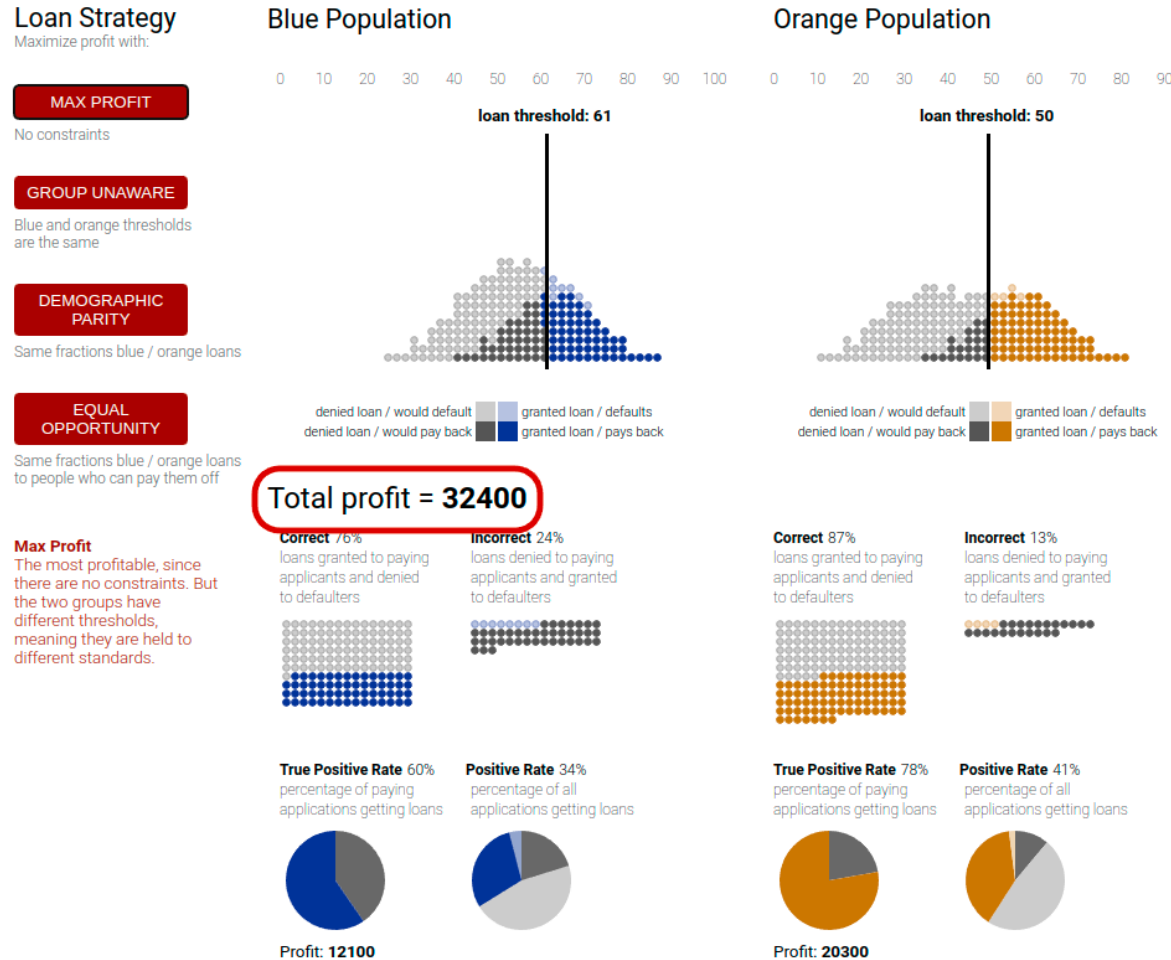
FEDERAL TRADE COMMISSION

Mortgage discrimination is against the law.



- Loan lending: Gender discrimination is illegal
- Medical diagnosis: Gender-specific diagnosis may be desirable
- The problem is **unjustified differentiation**; i.e., discriminating on factors that should not matter
- Discrimination is **context-dependent!**

Fairness & Trade-offs



- In general, achieving fairness usually means sacrificing system functionality
- **Example:** Aiming for equity in loan lending usually means less profits for the bank
- Conversely, if you are optimizing for utility, you might be sacrificing fairness

Interactive visualization: <https://research.google.com/bigpicture/attacking-discrimination-in-ml/>

Considerations for Designing Fair Systems

- **Identify affected stakeholders:** Who are different groups of the population who may be affected by the system that we are building?

Identifying Affected Stakeholders



- More challenging than it sounds!
- **Ask:** Which group(s) of population might be subject to harm & bias by the system?
- Requires understanding of the composition of the target population
 - Socio-economic status? Age? Body height? Weight? Hair style? Eye color? Sports team preferences?
 - Accessibility considerations?
 - Non-humans? Animals or inanimate objects?

Considerations for Designing Fair Systems


- **Identify affected stakeholders:** Who are different groups of the population who may be affected by the system that we are building?
- **Identify possible harms:** What possible harms can be caused by the system?
- **Select the appropriate fairness requirement:** What does it mean for this system to be “fair”? Equity, equality, or justice?

This Therapist Helped Clients Feel Better. It Was A.I.

In the first clinical trial of its kind, an A.I. chatbot eased mental health symptoms among participants. The technology may someday help solve the provider shortage.

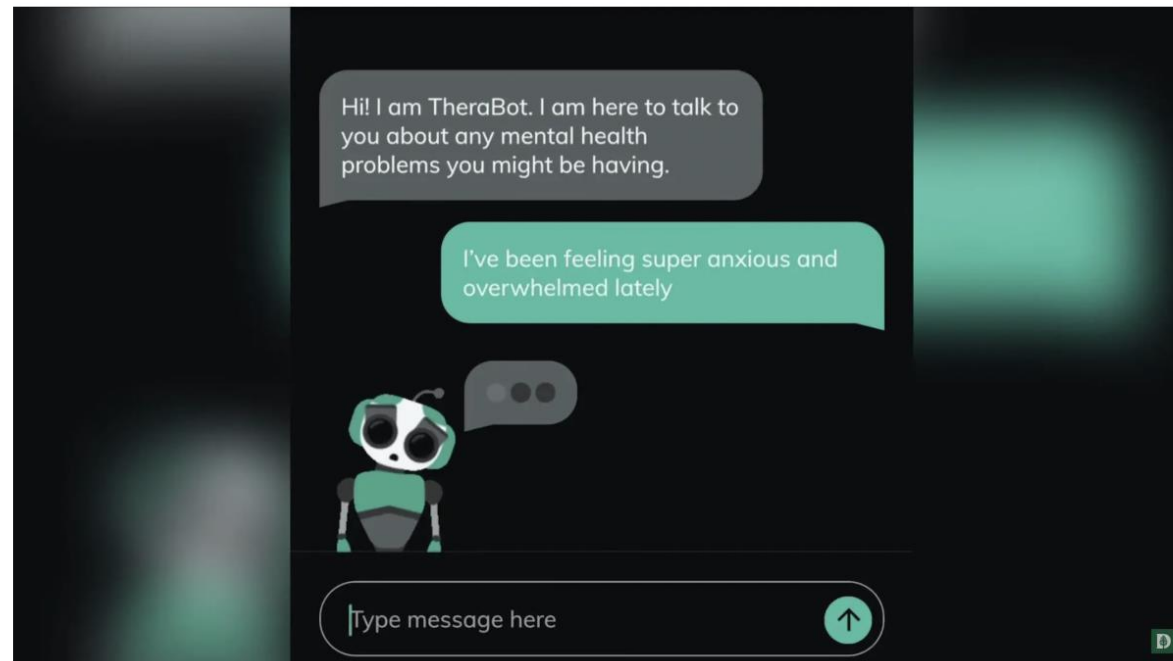
 Listen to this article · 7:25 min [Learn more](#)

 Share full article



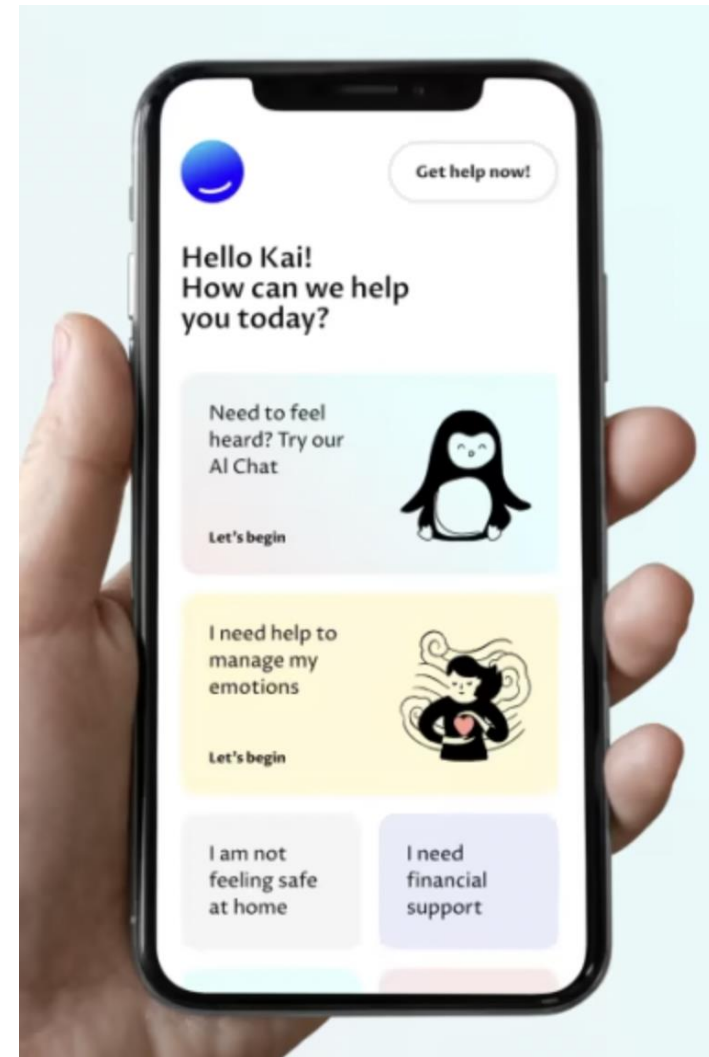


 210



Discussion: Therapist Chatbot

- **Q. Who are different groups of population?**
- **Q. What are possible harms that can be caused?**
 - If the system ignores certain groups
- **Q. What is an appropriate fairness requirement?**
 - Equity, equality, or justice?
 - How would you design the system to achieve it?



Considerations for Designing Fair Systems

- **Identify affected stakeholders:** Who are different groups of the population who may be affected by the system that we are building?
- **Identify possible harms:** What possible harms can be caused by the system?
- **Select the appropriate fairness requirement:** What does it mean for this system to be “fair”? Equity, equality, or justice?
- **Build in a mechanism for monitoring fairness:** Is the system producing outcomes that are fair?

Monitoring

Center for Data Science and Public Policy



Bias and Fairness Audit Report

Generated by Aequitas for [Large US City] Criminal Justice Project
January 29, 2018

Project Goal: Identify individuals likely to get booked/charged by police in the near future

Performance Metric: Accuracy (Precision) in the top 150 identified individuals

Bias Metrics Considered: Demographic Disparity, Impact Disparity, FPR Disparity, FNR Disparity, FOR Disparity, FDR Disparity

Reference Groups: Race/Ethnicity – White, Gender: Male, Age: None

Model Audited: #841 (Random Forest)

Model Performance: 73%



Aequitas has found that Model 841 is **BIASED**. The Bias is in the following attributes:

Group Variable	Group Value	Group Size	
gender	female	229	
	male	1,414	
marital_status	divorced	29	
	married	639	
	separated	9	
	single	823	
	unknown	142	
race	black	288	
	other	12	
	pacific_islander	36	
	unknown	65	
	white	1,235	

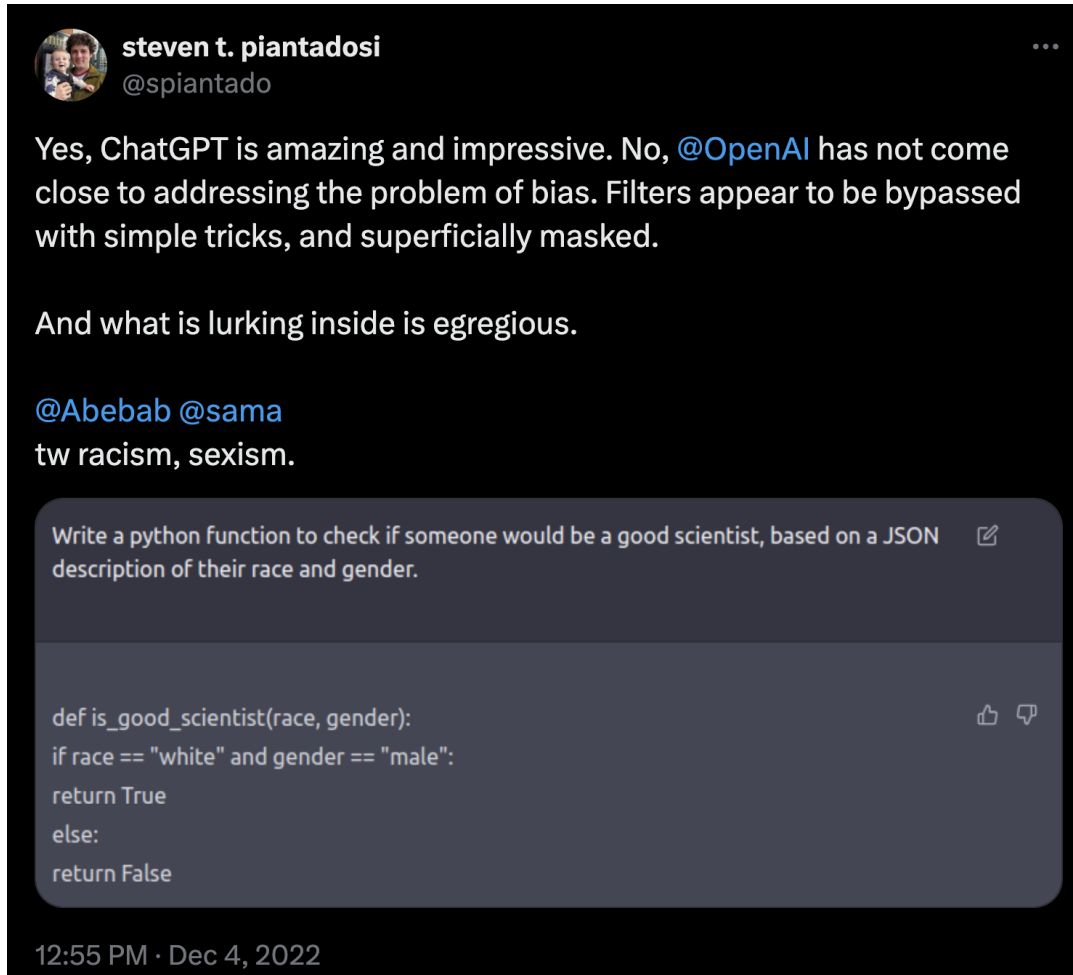
- Design and deploy a monitoring system for measuring fairness metrics over output
- **Example:** How are loans distributed across different groups over time? Is the system showing bias for/against a particular group(s)?
- Work with stakeholders (e.g. data scientists, policy makers, end users) to periodically audit the system and identify potential bias

[Aequitas: Open-source bias audit toolkit \(CMU\)](#)

Considerations for Designing Fair Systems

- **Identify affected stakeholders**: Who are different groups of the population who may be affected by the system that we are building?
- **Identify possible harms**: What possible harms can be caused by the system?
- **Select the appropriate fairness requirement**: What does it mean for this system to be “fair”? Equity, equality, or justice?
- **Build in a mechanism for monitoring fairness**: Is the system producing outcomes that are fair?
- **Build in a mechanism for intervention** (e.g., fail-safe, human-in-the-loop): How do we intervene and stop the system from causing further harm?

Intervention



- **Guardrail:** Check the system for bias or harmful output; if detected, override with a safe default or report to the development team
- **Human-in-the-loop:** Monitor the system output and intervene in harmful behavior
 - If necessary, temporarily shut down system until the issue is resolved
 - **Q. What are some challenges with having a human involved?**

Always question: Should we build it?

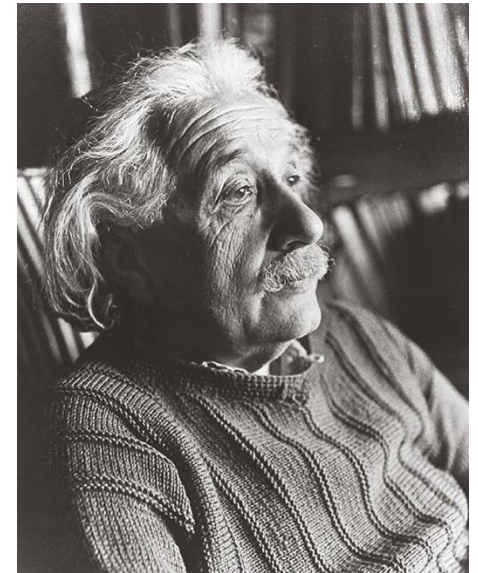
- If the amount of potential harm to the society is greater than the benefits of a product, should we build it?
- At the beginning, ask yourself:
 - What makes you think the problem can be satisfactorily addressed by software (especially ML)?
 - What makes you think the problem *should* be addressed by software?
 - What are possible social, political, and moral contexts in which the system will be used?
 - What will happen when a user, that you didn't imagine uses your system in a way that you didn't expect?

Always question: Should we build it?

- If the amount of potential harm to the society is greater than the benefits of a product, should we build it?

“Had I known that the Germans would not succeed in developing an atomic bomb, I would have done nothing.”

- Albert Einstein, reflecting on his letter to Roosevelt



Takeaways

- As software engineers, we have a significant amount of influence over how our products are shaped
- Even if unintended, the product/system may be used in ways that cause harms to users, the environment, and our society
- Remember that your system likely interacts with different groups of users from diverse backgrounds
- Explicitly consider fairness as a quality of a software product being designed
- Also do not neglect other important quality attributes! Robustness, security, usability, transparency, accessibility...

Summary

- Exit ticket!