CS 4530
Fundamentals of Software Engineering

Module 17: Engineering Software for Equity

Adeel Bhutta, 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 human values in designing software systems
  - Explain some techniques that software engineers can use in producing software systems that are more congruent with human values.

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…

A few representative examples

- Some of these are old, but things haven't changed…

Following regulations alone is not enough

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

Algorithmic sentencing can discriminate

- 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 can discriminate

- …against the poorest of us

THE WALL STREET JOURNAL.

Websites Vary Prices, Deals Based on Users' Information

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.

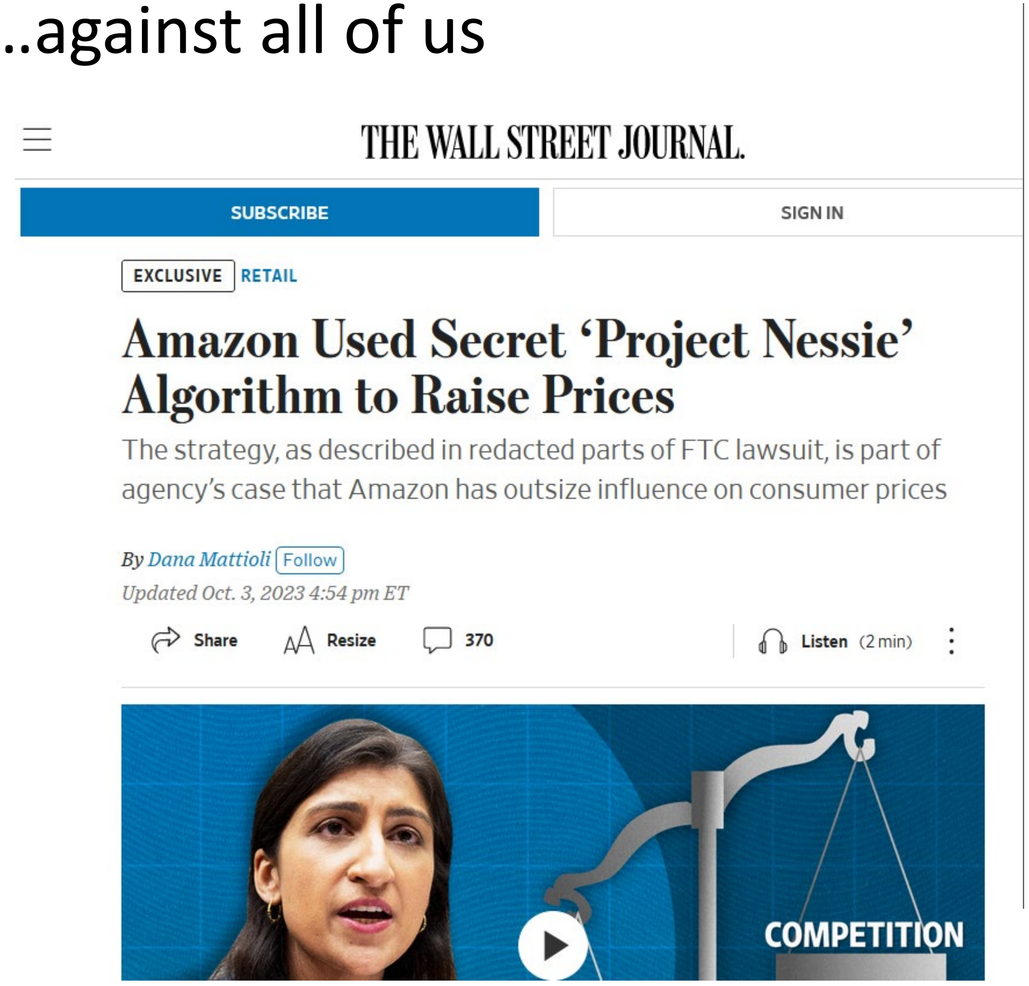
We instantiate the UA framework in *FairTest*, the first comprehensive tool that helps developers check data-driven applications for unfair user treatment. It enables scalable and statistically rigorous investigation of associations between application outcomes (such as prices or premiums) and sensitive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

Algorithms can be used for price gouging

- …against all of us



THE WALL STREET JOURNAL.

SUBSCRIBE SIGN IN

EXCLUSIVE RETAIL

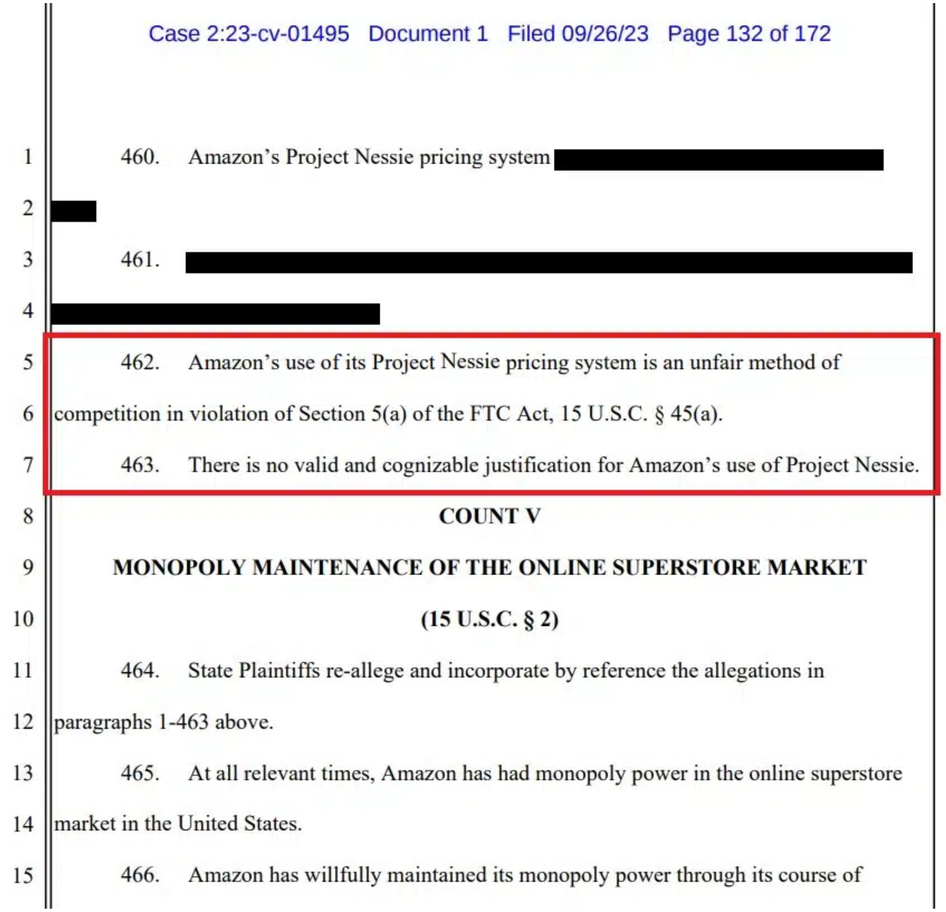
Amazon Used Secret 'Project Nessie' Algorithm to Raise Prices

The strategy, as described in redacted parts of FTC lawsuit, is part of agency's case that Amazon has outsize influence on consumer prices

By Dana Mattioli Follow

Updated Oct. 3, 2023 4:54 pm ET

Share Resize 370 Listen (2 min)



Case 2:23-cv-01495 Document 1 Filed 09/26/23 Page 132 of 172

460. Amazon's Project Nessie pricing system [redacted]

[redacted]

461. [redacted]

[redacted]

462. Amazon's use of its Project Nessie pricing system is an unfair method of competition in violation of Section 5(a) of the FTC Act, 15 U.S.C. § 45(a).

463. There is no valid and cognizable justification for Amazon's use of Project Nessie.

COUNT V

MONOPOLY MAINTENANCE OF THE ONLINE SUPERSTORE MARKET

(15 U.S.C. § 2)

464. State Plaintiffs re-allege and incorporate by reference the allegations in paragraphs 1-463 above.

465. At all relevant times, Amazon has had monopoly power in the online superstore market in the United States.

466. Amazon has willfully maintained its monopoly power through its course of

https://www.youtube.com/watch?v=T1Bcupz77-Q

https://www.wsj.com/business/retail/amazon-used-secret-project-nessie-algorithm-to-raise-prices-6c593706

UIs can discriminate against the differently-abled

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

Pizza LLC v. Robles

"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.**

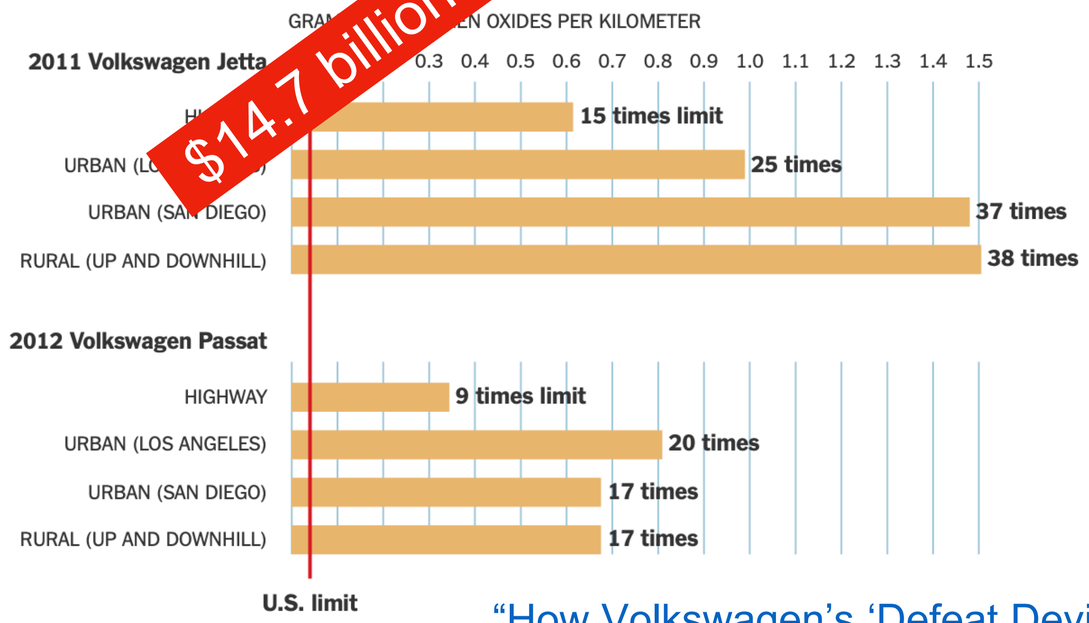
Software can help to evade regulation

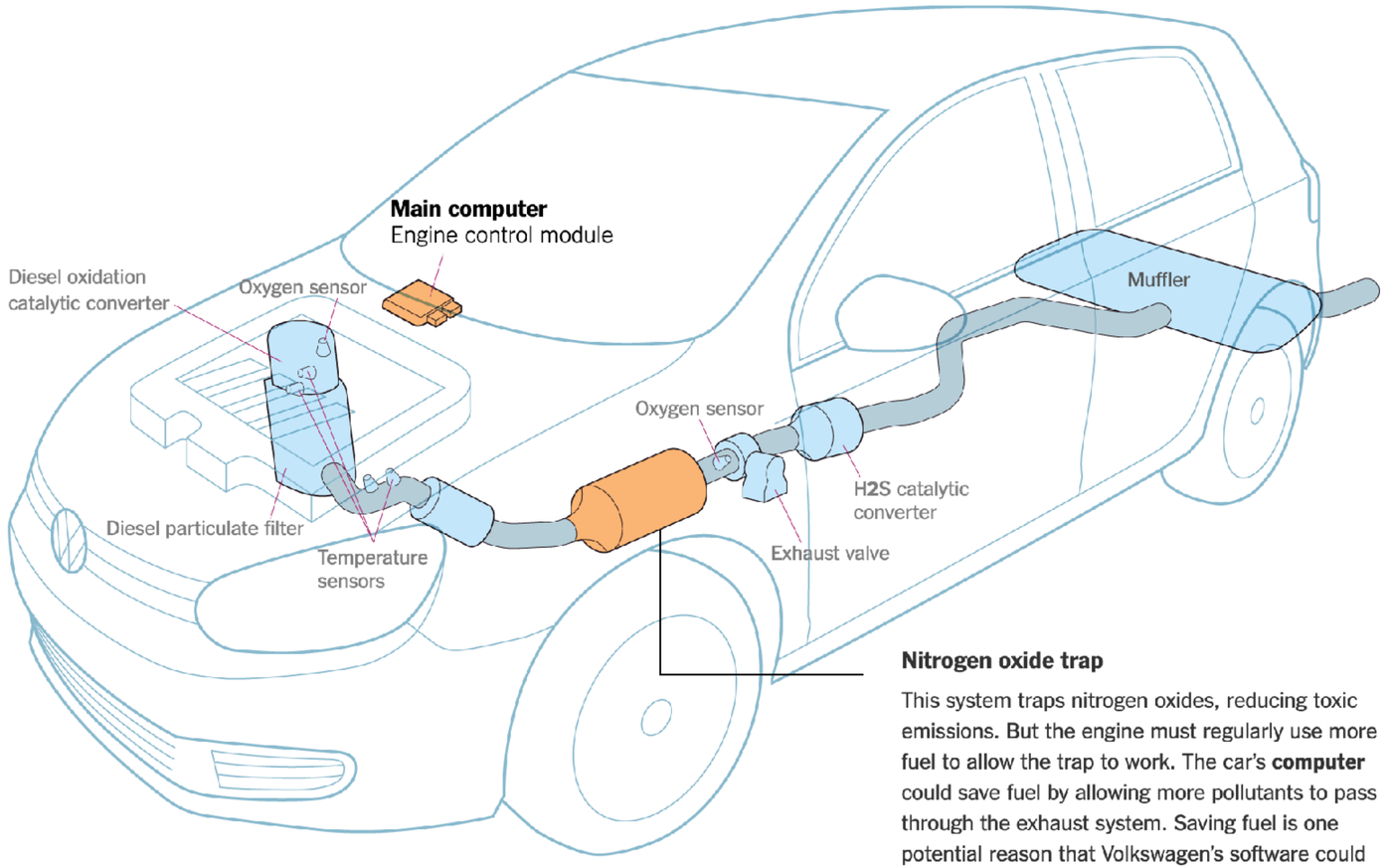
The Emissions Tests That Led to the Discovery of VW's Cheating

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… …trogen oxides.

$14.7 billion settlement

Average emissions of nitrogen ox… …-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

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

Training Data can be biased (and usually is)

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/

Training of AI systems can impact 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

AI programs can hallucinate

Lawyer cites fake cases generated by ChatGPT in legal brief

The high-profile incident in a federal case highlights the need for lawyers to verify the legal insights generated by AI-powered tools.

Published May 30, 2023

The ChatGPT Lawyer Explains Himself

In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he "did not comprehend" that the chat bot could lead him astray.

Here's What Happens When Your Lawyer Uses ChatGPT

A lawyer representing a man who sued an airline relied on artificial intelligence to help prepare a court filing. It did not go well.

scientific reports

Explore content | About the journal | Publish with us

nature > scientific reports > articles > article

Article | Open access | Published: 07 September 2023

Fabrication and errors in the bibliographic citations generated by ChatGPT

William H. Walters & Esther Isabelle Wilder

Scientific Reports **13**, Article number: 14045 (2023) | Cite this article

5866 Accesses | **2** Citations | **95** Altmetric | Metrics

ChatGPT is making up fake Guardian articles. Here's how we're responding

Chris Moran

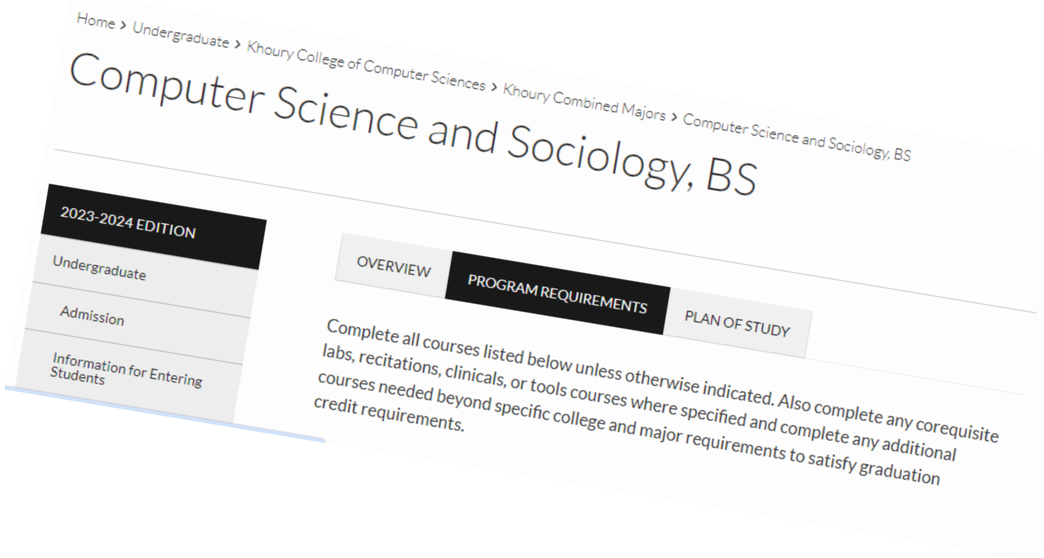
And this is only the tip of the iceberg

- Other Challenges:
  - interfaces and systems designed to be addictive;
  - corporate ownership of personal data;
  - weak cyber security and personally identifiable information (PII) protection;
  - and many more …

SOCL 4528. Computers and Society. (4 Hours)

Focuses on the social and political context of technological change and development. Through readings, course assignments, and class discussions, offers students an opportunity to learn to analyze the ways that the internet, artificial intelligence, and other technological advances have required a reworking of every human institution—both to facilitate the development of these technologies and in response to their adoption.

Attribute(s): NUpath Difference/Diversity, NUpath Societies/Institutions



Home › Undergraduate › Khoury College of Computer Sciences › Khoury Combined Majors › Computer Science and Sociology, BS

Computer Science and Sociology, BS

2023-2024 EDITION

Undergraduate

Admission

Information for Entering Students

OVERVIEW | PROGRAM REQUIREMENTS | PLAN OF STUDY

Complete all courses listed below unless otherwise indicated. Also complete any corequisite labs, recitations, clinicals, or tools courses where specified and complete any additional courses needed beyond specific college and major requirements to satisfy graduation credit requirements.

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 weigh competing interests and values
  - 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*)

Technological Success takes a broader view

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.

Maybe we should 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

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?

Identifying and Filtering 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…

Not every impacted individual has **legitimate values** at play. Those stakeholders 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

Ownership and property

Privacy

Trust

Accountability

Addressing Value Tensions

This is where the hard choices happen

What are the core values that cannot be violated?

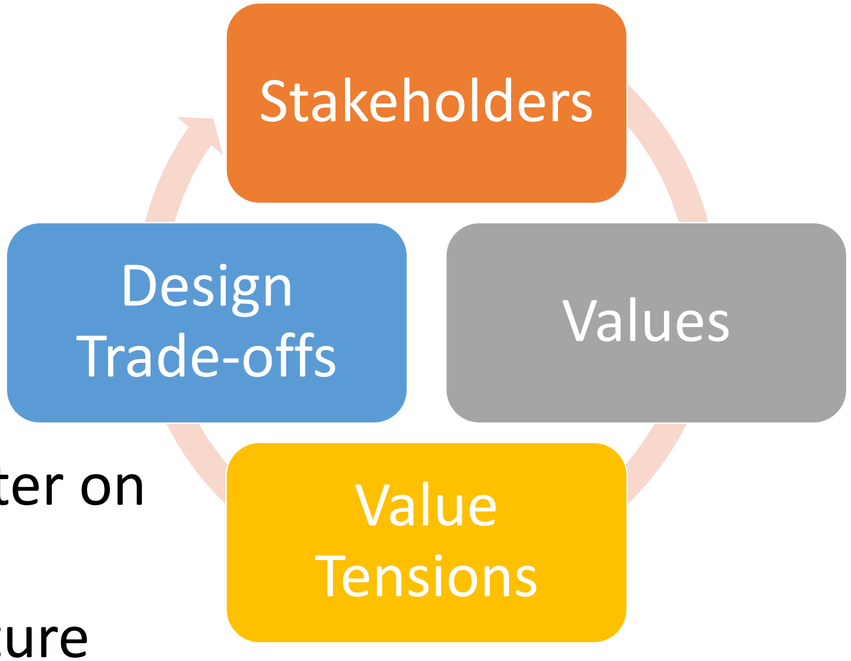
Which tensions can be addressed through:

- 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?

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

Example 1: 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

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.

Review

- You should now be able to…
  - Illustrate how software can cause inadvertent harm or amplify inequities
  - Explain the role of human values in designing software systems
  - Explain some techniques that software engineers can use in producing software systems that are more congruent with human values.