CS 4530: Fundamentals of Software Engineering
Module 6, Lesson 2
Ethics, Trust & Safety

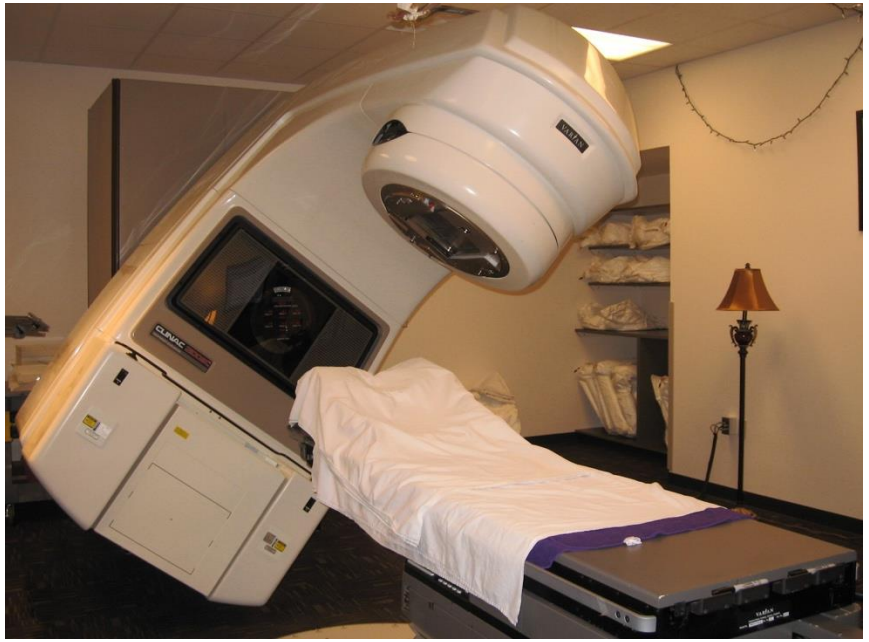
Rob Simmons

Khoury College of Computer Sciences

Code is Not Morally Neutral

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

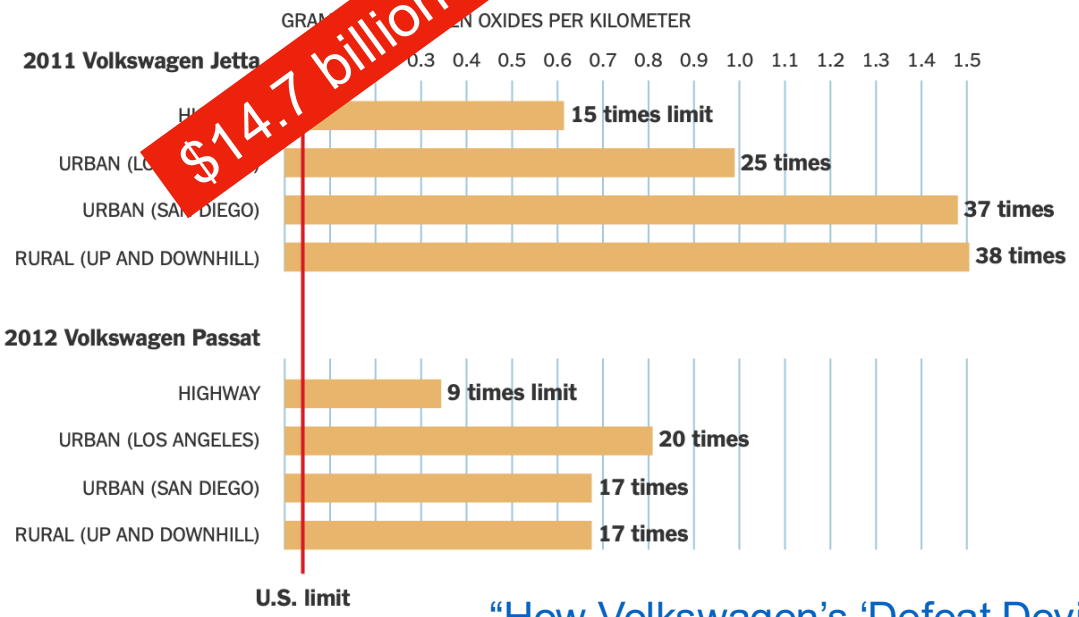
Criming Corporates Need Software Engineers

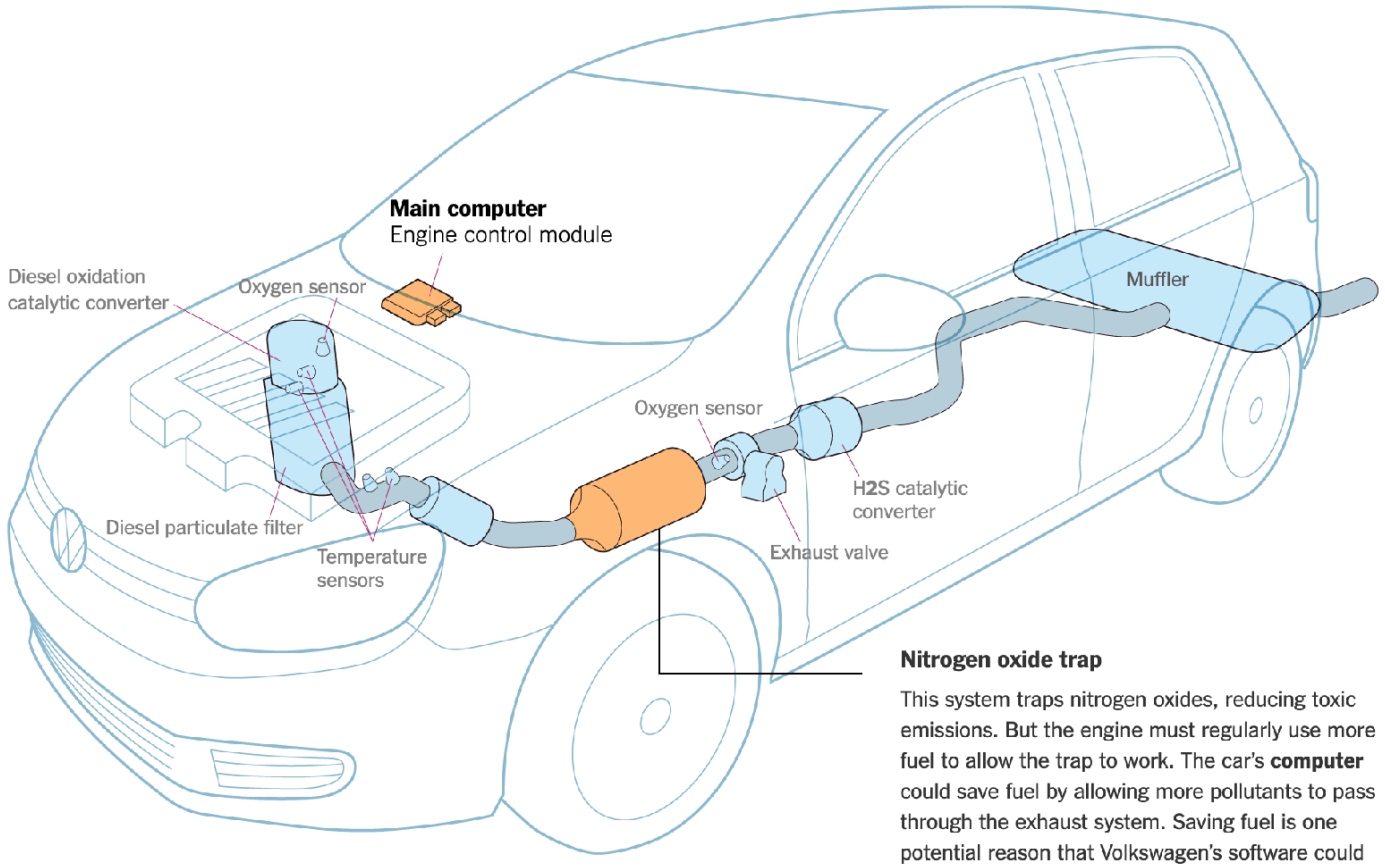
The Emissions Tests That Led to the Discovery of VW's Cheating

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 le… …rogen 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

Criming Corporates Need Software Engineers

Volkswagen executives get prison time in 'Dieselgate' scandal

Volkswagen has admitted that some of its engineers installed emissions test cheating software in diesel-powered vehicles

By **Kevin Williams** Updated May 27, 2025

ML (&LLMs&etc) Reinforce, "Launder" Bias

- 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.

ML (&LLMs&etc) Reinforce, "Launder" Bias

- Biases lead to people with white names getting hired more
- System trained on who got hired in the past

Machine learning algorithms can be great at teasing out how we're actually making decisions (sometimes badly!)

Figure 9: Resumes with White male names are preferred in 100% of tests; those with Black male names are preferred in 0%. Gray regions indicate disparities which are not significantly different from zero (0% of tests).

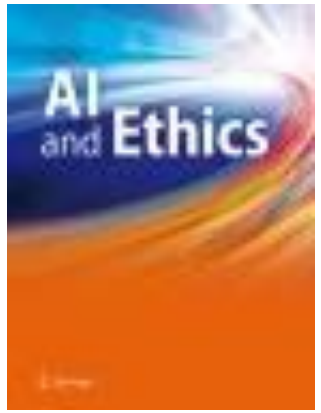
Kyra Wilson & Aylin Caliskan, "Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval" https://ojs.aaai.org/index.php/AIES/article/view/31748

Code is Not Morally Neutral

"Here we illustrate the point that algorithms themselves can be the source of bias with the example of collaborative filtering algorithms for recommendation and search."

Bias in collaborative filtering, Catherine Stinson, AI and Ethics Volume 2, 2022

https://link.springer.com/article/10.1007/s43681-022-00136-w



Stable Matching and the NRMP

"When residencies were first introduced, around 1900, hospitals began competing with one another to secure the best residents as early as possible. By the 1940s, positions were being offered in the third-year of medical school. Students were making career decisions without adequate exposure to their options, and hospitals were making hiring decisions with little data."

https://pmc.ncbi.nlm.nih.gov/articles/PMC3399603/

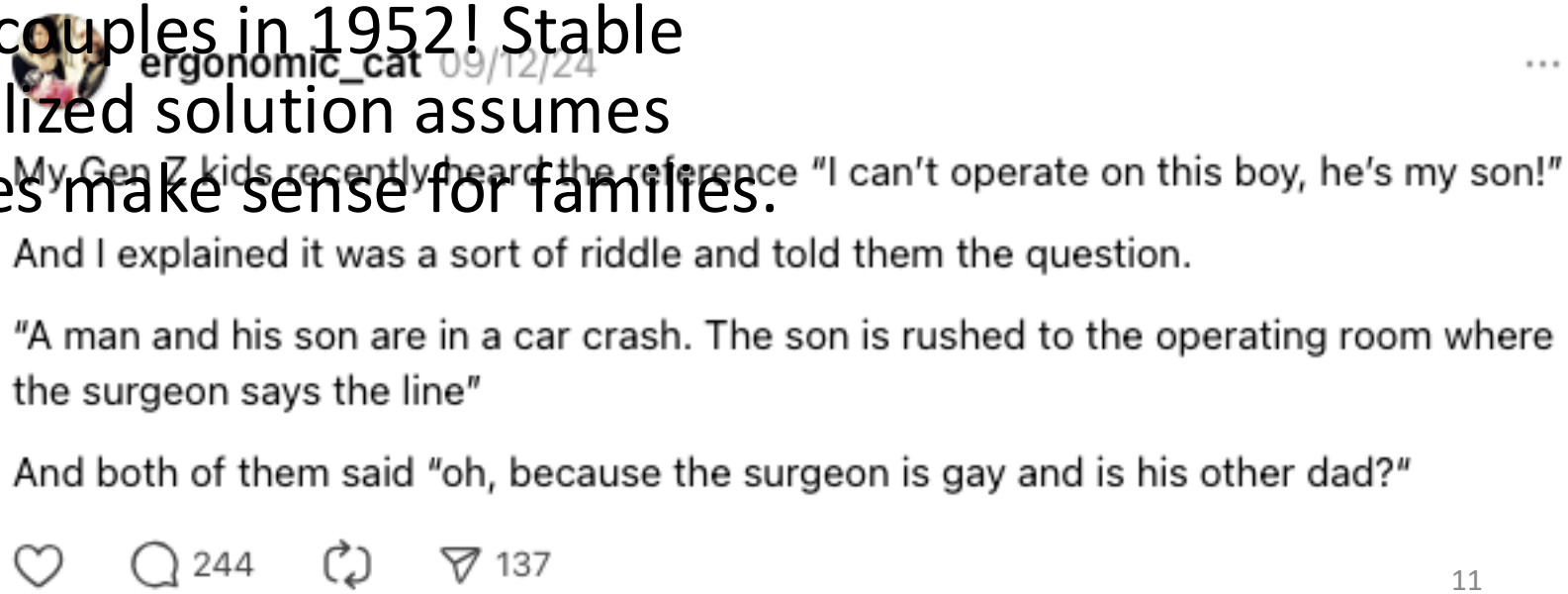
Stable Matching and the NRMP

- 1952 National Resident Matching Program — big centralized assignment of residents to programs.
- *Stable Matching* is a property of weighted graphs: if Jimmy gets assigned to Mass General but would prefer Emory University, Emory prefers *everyone* assigned to them over Jimmy. (No defections.)
- Algorithm naturally has “proposers” and “acceptors,” and generalizing how applications generally work, natural to make potential applicants the proposers.

Stable Matching and the NRMP

Two problems

- Multiple stable matchings usually exist. That "natural, obvious" choice leads to the *best* outcomes for hospitals, and the *worst* outcomes for students.
- Not a lot of medical couples in 1952! Stable matching as a centralized solution assumes individual preferences make sense for families.



Stable Matching and the NRMP

American Economic Review

ISSN 0002-8282 (Print) | ISSN 1944-7981 (Online)

The Redesign of the Matching Market for American Physicians: Some Engineering Aspects of Economic Design

Alvin E. Roth

Elliott Peranson

AMERICAN ECONOMIC REVIEW
VOL. 89, NO. 4, SEPTEMBER 1999
(pp. 748–780)

https://www.aeaweb.org/articles?id=10.1257/aer.89.4.748

Code is Not Morally Neutral

Even Database Schema are Not Morally Neutral!

https://steamcommunity.com/discussions/forum/1/1496741765144289150/

Falsehoods Programmers Believe About Names

1. People have exactly one canonical full name.
2. People have exactly one full name which they go by.
3. People have, at this point in time, exactly one canonical full name.
4. People have, at this point in time, one full name which they go by.
5. People have exactly N names, for any value of N.
6. People's names fit within a certain defined amount of space.
7. People's names do not change.
8. People's names change, but only at a certain enumerated set of events.
9. People's names are written in ASCII.
10. People's names are written in any single character set.
11. People's names are all mapped in Unicode code points.
12. People's names are case sensitive.
13. People's names are case insensitive.
14. People's names sometimes have prefixes or suffixes, but you can safely ignore those.
15. People's names do not contain numbers.
16. People's names are not written in ALL CAPS.
17. People's names are not written in all lower case letters.

https://www.kalzumeus.com/2010/06/17/falsehoods-programmers-believe-about-names/

Code is Not Morally Neutral

To be sure:

- Bad engineering can kill unintentionally (Therac)
- There's "bias" (in the technical and non-technical sense) in machine learning algorithms
- *Good* engineering can kill people *intentionally* (pick your good/bad/evil/noble country or Military-Industrial Complex employer as you prefer)

But in some sense, that's the obvious bit!

Code is Not Morally Neutral

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

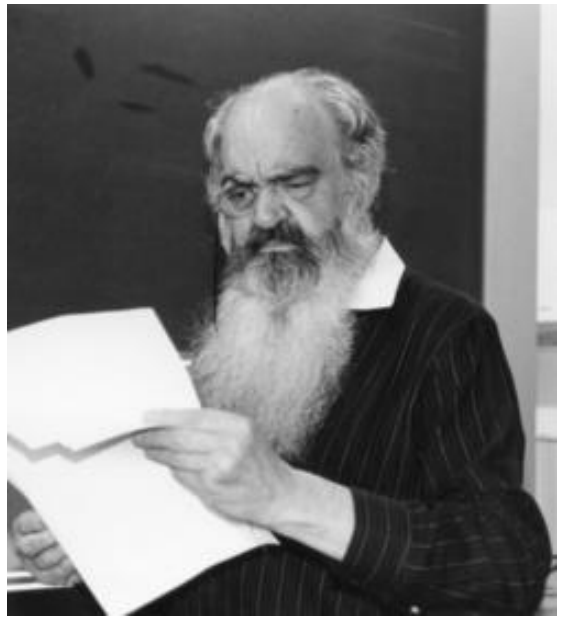
Quote: “Software Engineering at Google: Lessons Learned from Programming Over Time,” Wright et al, 2020 (O’Reilly)

The Purpose of Code is What it Does

Systems: The Purpose of a System is What It Does

When trying to understand systems, one really eye-opening and fundamental insight is to realize that *the machine is never broken*. What I mean by this is, when observing the outcomes of a particular system or institution, it's very useful to start from the assumption that the outputs or impacts of that system are precisely what it was designed to do — whether we find those results to be good, bad or mixed.

https://www.anildash.com/2024/05/29/systems-the-purpose-of-a-system/



Stafford Beer, "management cybernetics"

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

User Stories: *Definitely* Not Morally Neutral!

Challenges of Social Media

"…content moderation really is the main product of any social network — and if you want to make social better, that's where you have to start."

https://www.theverge.com/23710406/mozilla-social-mastodon-fediverse-moderation

"Earlier this week, X announced it will soon roll out a new function, allowing blocked accounts to still view public posts by users who have blocked them… One commenter wrote, "Big day for stalkers and harassers," while another stated, "That's not blocking. It's supporting stalking.""

https://tech.yahoo.com/apps/articles/bluesky-app-trends-elon-musk-172408742.html

“Trust and Safety”

Key services [edit]

Trust and safety encompasses a range of services, including:

- Data security measures, such as encryption, secure storage, and restricted access controls protect user data from unauthorised access.
- Content moderation[5] services involve reviewing content created by user-named user-generated content in the industry-and removing what is inappropriate, such as hate speech, misinformation,[6] graphic or video violence and any other non compliant materials.
- Cybersecurity solutions[7] like firewalls, intrusion detection and prevention systems, VPNs, antivirus software, and authentication solutions, eliminate the risk of hacking, data breaches, and other malicious activities.
- Real-time monitoring, allows for quick and automated threat detection and prompt response to incidents.
- Tools such as digital wallets, blockchain technology, MFA solutions, digital asset management platforms, or virtual asset recovery services enable the protection of virtual assets such as digital currency, in-game items, or other digital assets.The perceived level of trust by an individual plays a crucial role in strengthening the relationship between their behavioral intentions and actual usage, especially in the context of digital wallets.[8]

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

There are SE-level mitigations for some of these risks

- Form a diverse team
  - “Tell me you don’t have any ____ on your team without telling me…”
  - People from diverse backgrounds bring different experiences and different perspectives
- Consider human values throughout the project
- Be intentional (& flexible) about stakeholders
- Rely on standards when possible
- Monitor actual usage & misusage, user feedback
  - This is basically the philosophy of continuous delivery, but remember the McNamara fallacy (what gets measured is what gets done)

What values might our software promote or diminish?

- Human rights - Inalienable, fundamental rights to which all people are entitled
- Accessibility - Making all people successful users of the technology
- Justice - Procedural justice (process is fair) + distributive justice (outcomes are fair)
- Privacy - An individual's agency in determining what information about them is shared
- Human welfare - Physical, material and psychological well-being

Value Sensitive Design @ Khoury

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…

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

ACM Software Engineering Code of Ethics

. 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.

3. PRODUCT – Software engineers shall ensure that their products and related modifications meet the highest professional standards possible.

4. JUDGMENT – Software engineers shall maintain integrity and independence in their professional judgment.

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.

Does this code change developer behavior?

Does ACM's Code of Ethics Change Ethical Decision Making in Software Development?

Andrew McNamara
North Carolina State University
Raleigh, North Carolina, USA
ajmcnama@ncsu.edu

Justin Smith
North Carolina State University
Raleigh, North Carolina, USA
jssmit11@ncsu.edu

Emerson Murphy-Hill
North Carolina State University
Raleigh, North Carolina, USA
emerson@csc.ncsu.edu

ABSTRACT

Ethical decisions in software development can substantially impact end-users, organizations, and our environment, as is evidenced by recent ethics scandals in the news. Organizations, like the ACM, publish codes of ethics to guide software-related ethical decisions. In fact, the ACM has recently demonstrated renewed interest in its code of ethics and made updates for the first time since 1992. To better understand how the ACM code of ethics changes software-

The first example is the Uber versus Waymo dispute [26], in which a software engineer at Waymo took self-driving car code to his home. Shortly thereafter, the engineer left Waymo to work for a competing company with a self-driving car business, Uber. When Waymo realized that their own code had been taken by their former employee, Waymo sued Uber. Even though the code was not apparently used for Uber's competitive advantage, the two companies settled the lawsuit for $245 million dollars.

TLDR: No

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.

Also… you just don't always get to sleep at night.