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 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…

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*)

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?

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

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

This is where the hard choices happen

What are the core values that cannot be violated?

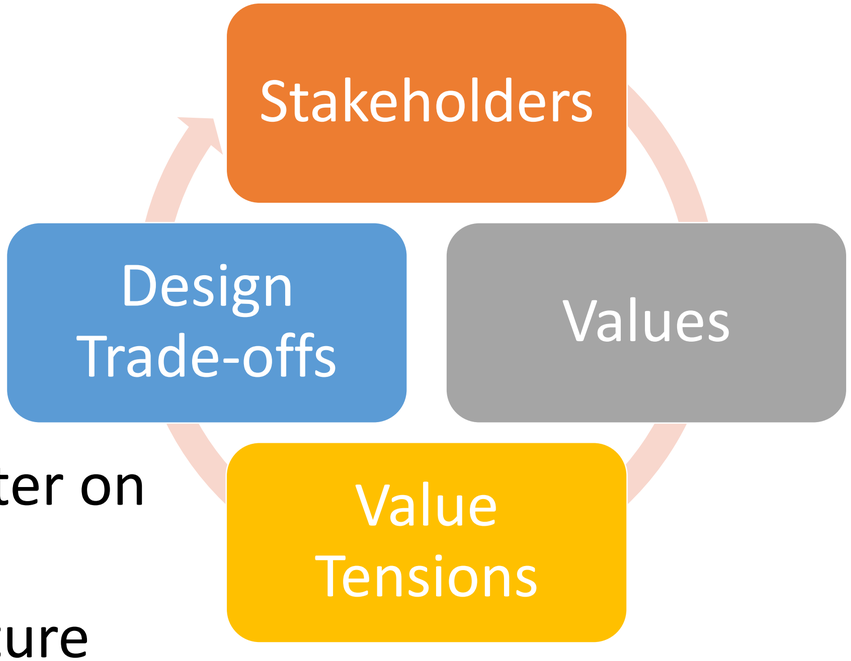
Which tensions can be addressed through:

- 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 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.

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.

After this point are the old slides