INTRO TO VALUE SENSITIVE DESIGN

Khoury College of Computer Sciences

Northeastern University

John Basl, Ron Sandler, Christo Wilson

Edited 10/21/19

VALUES AND TECHNOLOGY

- 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

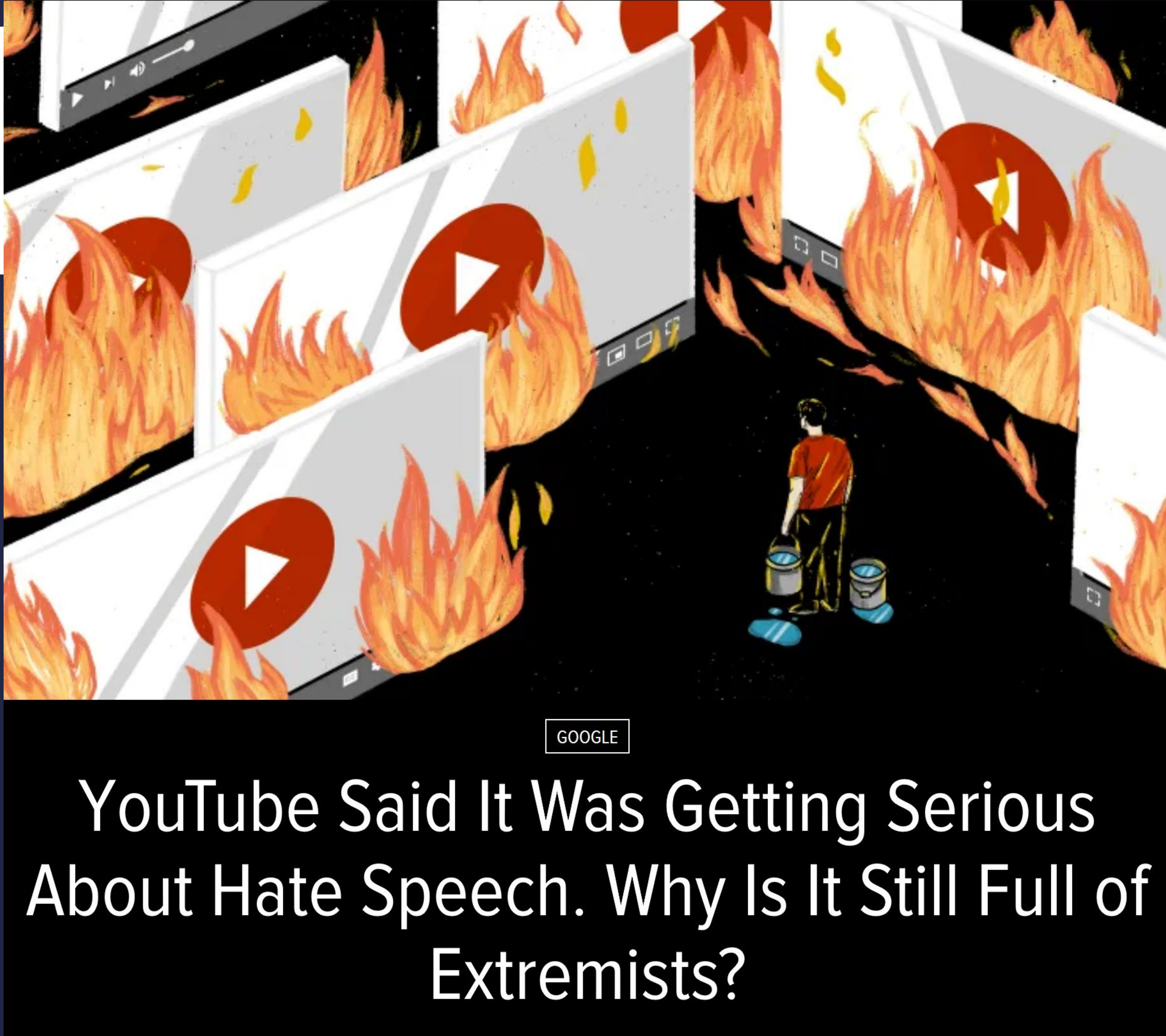
MOTIVATING EXAMPLE: CONTENT MODERATION

Platforms take a “neutral” approach to moderating speech

- First amendment
- CDA 230

Real world consequences

- Violent radicalization
- Genocide in Myanmar



GOOGLE

YouTube Said It Was Getting Serious About Hate Speech. Why Is It Still Full of Extremists?

BAD SOLUTIONS

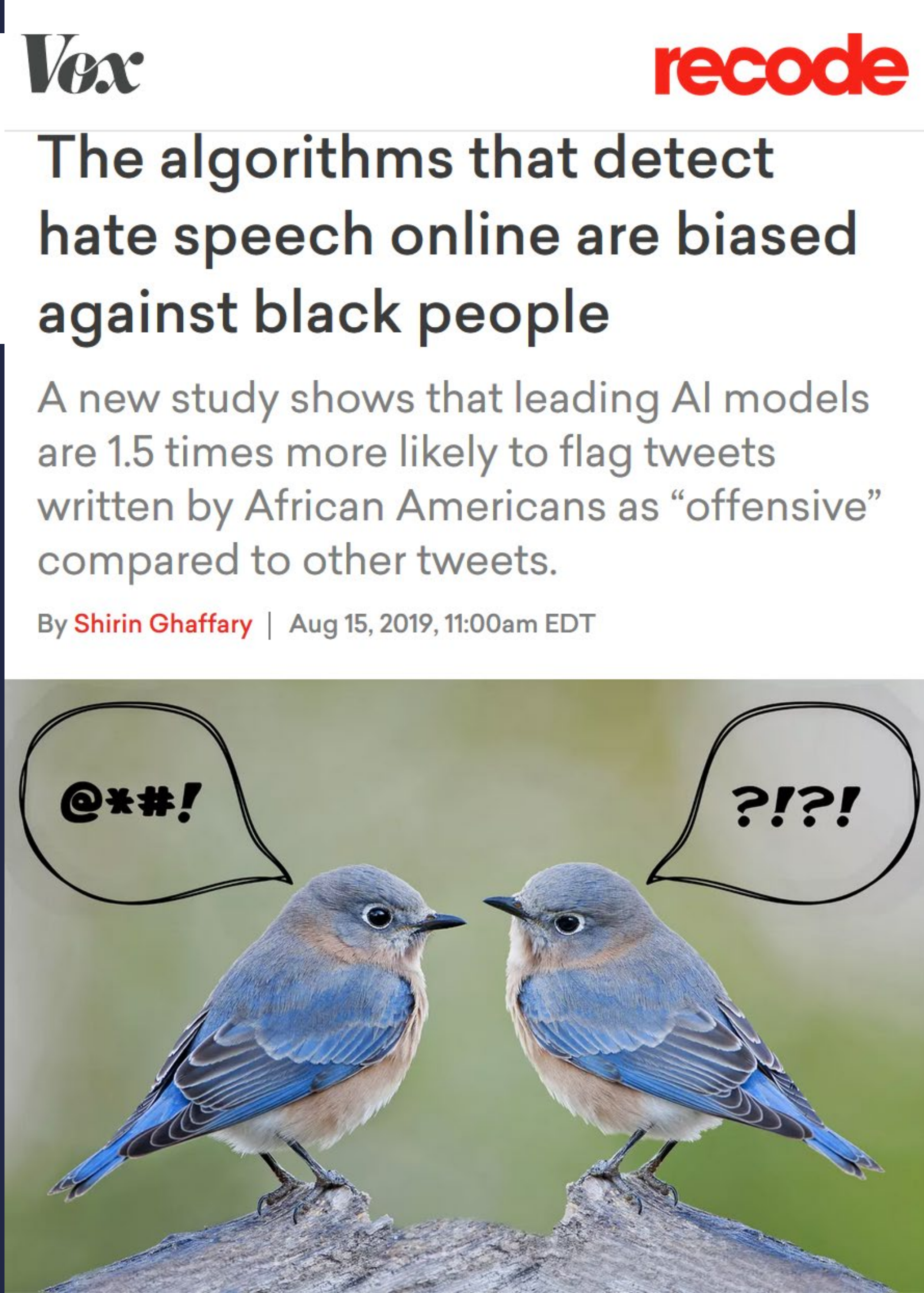
Idea: train a machine learning model to detect hate speech

- Fast and scalable
- “Fair” because it doesn’t rely on human judgement

Harms the people it was supposed to protect

Fails to take context into account

- Historical quotations vs. expressions of opinion
- Ownership of slurs by people from targeted groups



Vox recode

The algorithms that detect hate speech online are biased against black people

A new study shows that leading AI models are 1.5 times more likely to flag tweets written by African Americans as “offensive” compared to other tweets.

By Shirin Ghaffary | Aug 15, 2019, 11:00am EDT

CONTENT MODERATION DONE RIGHT?

Defining hate speech

- Examining socio-historical power relations between groups
- Soliciting diverse perspectives on the problem and possible solutions

Taking context into account

- Historical quotes, artistic expression, in-group vs. out-group speaker…

Implementing a process

- Crowdsourced detection? Automation with AI? Paid content moderators?
- Process for contesting decisions and requesting a review?

Deciding how to enforce the policy

- Censor content and/or ban the speaker?
- Flag content and show a warning label or collapse it (i.e. require a click to view)?

Taking the delivery medium into account

- Should hate speech appear in search results?
- Should hate speech be algorithmically promoted or recommended (e.g. in a news feed)?

Answering all of these questions require making challenging, normative choices!

MORE EXAMPLES

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

Facial recognition

Smart speakers that secretly have humans transcribe the recordings

Ridesharing apps that congest streets; home sharing apps that displace residents

Third-party data collection for hyper-targeted advertising

Self driving cars, autonomous drones

AIs that gate access to loans, insurance, employment, government services…

Deceptive user interface "dark patterns"

And on… and on… and on…

VALUES AND TECHNOLOGY

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*)

VALUE SENSITIVE DESIGN (VSD)

The goal of Value Sensitive Design is to make socially-informed and thoughtful value-based choices in the technology design process

1. Appreciating that technology design is a value-laden practice
2. Recognizing the value-relevant choice points in the design process
3. Identifying and analyzing the values at issue in particular design choices
4. Reflecting on those values and how they can or should inform technology design

VSD IN BRIEF

VSD is a…

Outlook for seeing the values in technology design

Process for making value-based choices within design

- Combines empirical, value, and technical investigations
- Design solutions that incorporate the values held by stakeholders
- Considers problems and solutions from diverse perspectives

VSD is not…

A moral framework or system of ethics

- It does not tell you what decisions to make
- Rather, it incorporates value reflections into the choosing process

It does not provide an algorithm for making decisions

- *No easy answers*
- Takes sustained commitment

MODES OF INQUIRY

Empirical Investigation

- How do stakeholders prioritize competing values?
- Differences between *what people say* and *what people do*?
- Economic incentives?
- What are the benefits/costs and their distributions?

Value Investigation

- What is the overall goal of the technology?
- What is the social context in which the technology will be situated?
- What values are at stake?
- Which stakeholders are legitimately impacted?
- What value-oriented criteria will be used to gauge project success?

Technical Investigation

- What frameworks and tools enable designers to meet value-oriented goals?
- Impact of law, policy, and regulation on your design?
- What about cybersecurity?
- Do quantifiable objectives align with value-oriented criteria?

EXPERTISE IN INQUIRY

Empirical Investigation

Where?

- In the field, gathering knowledge about the world

Disciplinary skills

- Sociology
- Behavioral economics
- Experimental psychology
- Political science

Value Investigation

Where?

- In front of a whiteboard

Disciplinary skills

- Applied ethics
- Critical race and gender theory
- Law and policy
- Environmental analysis

Technical Investigation

Where?

- In the computer, analyzing data and prototyping

Disciplinary skills

- Computer science
- Data science
- Cybersecurity/privacy

TOOLS OF INQUIRY

Empirical Investigation

- Observational studies
- Surveys
- Semi-structured interviews
- Experimental manipulations (A/B testing)
- Collection of primary source documents

Value Investigation

- Discussion and iteration with a multi-disciplinary team
- Contextualization of empirical results
- Case studies

Technical Investigation

- User-centered investigation of prototypes
- Auditing data and algorithms for bias
- Red team cybersecurity audits
- Privacy impact assessments

VSD IN ACTION

1. **Framing Technical Work**
   - Clarify explicitly supported project values and designer stance
   - Situate the work within a social context
2. **Empirical Investigation**
   - Identify key direct and indirect stakeholders
   - Elicit potential values from stakeholders
   - Systematically identify benefits and harms for stakeholders
   - Refine the social context
3. **Conceptual Investigation**
   - Develop working definitions of key values and identify potential value tensions
   - Define technical and technological success objectives
   - Map tensions to success objectives
4. **Technical Investigation**
   - Identify choice points where the design team has the mandate, control, or power to intervene
   - Build technological and social interventions
5. **Monitor and Respond to Change Over Time**

CHALLENGES

Building ethical technology is not easy. That doesn’t mean we can ignore the 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 value tensions?
- 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 SOCIAL STRUCTURES

All technology is socially situated. What is the social structure around your technology?

What are the benefits and costs of the technology?

How are the benefits and costs distributed across individuals and society?

Are there socio-economic or historical inequalities?

How could the technology change the activity into which it is introduced?

How will social structures change over time?

Answering these questions requires empirical inquiry!

IDENTIFYING STAKEHOLDERS

Whose values are impacted by a piece of technology?

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…

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

What values are relevant to different, legitimate stakeholders?

Some values are universal: accessibility, justice, human rights, privacy

Others are tied to specific stakeholders and social contexts

Identifying implicated values

1. Grounded in a thorough conceptual and empirical understanding of the relevant features of the social situation
2. Informed by experience/knowledge from similar technologies or design decisions (case studies, etc.)
3. Refined through empirical investigation

Reflect on the scale of impacts to various stakeholders – focus on the major challenges

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

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

EXAMPLE VALUES

Autonomy refers to people’s ability to decide, plan, and act in ways that they believe will help them to achieve their goals

Informed consent refers to garnering people’s agreement, encompassing criteria of disclosure, comprehension, voluntariness, competence, and agreement

Identity refers to people’s understanding of who they are over time, embracing both continuity and discontinuity over time

Environmental sustainability refers to sustaining ecosystems such that they meet the needs of the present

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?

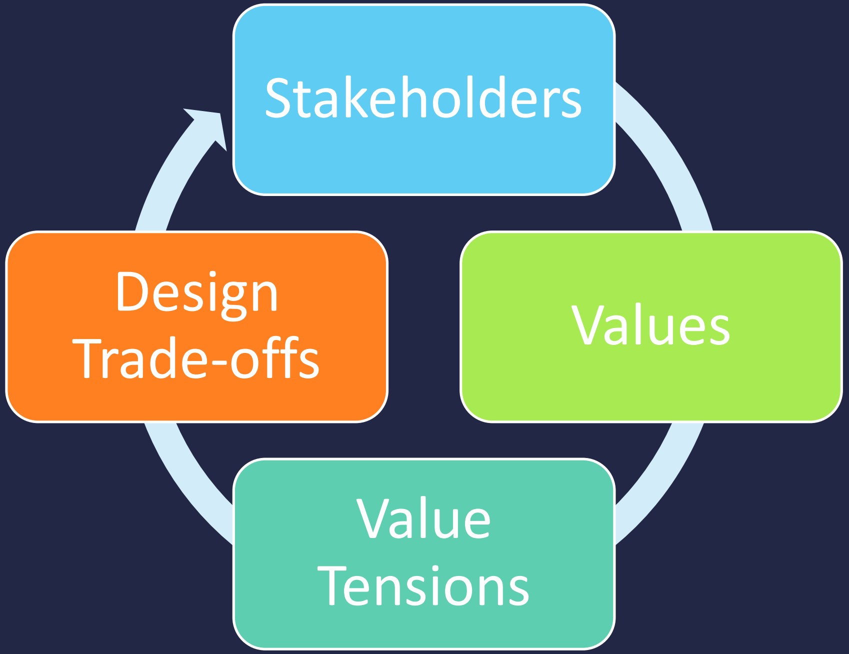
Which tensions can be addressed through:

- 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



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

EXAMPLE: “GOING DARK”

The issue: *rule of law* in tension with *security* and *privacy*

- Law enforcement wants access to data so they can identify and prosecute crimes
- End-to-end encryption prevents many avenues of data access for law enforcement
- But, encryption also preserves people’s digital privacy

Bad take: valuing *rule of law* above all else

- Requiring encryption back doors does give law enforcement access…
- But, back doors allow malicious parties to compromise encryption as well

Good take: framing the debate in terms of costs

- Individuals and companies rely on encryption to secure data, untold trillions of $ at stake
- Law enforcement’s interests are legitimate, but crime is relatively rare
- Bonus: law enforcement has other tools to access data besides cracking encryption

TIPS 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

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?

  Failure to consider incentive alignment
- Our database is breached and publicly released?

  Failure to consider security and privacy
- Our facial recognition AI is used to identify and harass peaceful protestors?
- Our child safety app is used to stalk women?

  Failure to consider appropriation across contexts and dual uses of technology
- Our chatbot is sexist or racist?

  Failure to consider biases in data

VSD IN ACTION

1. **Framing Technical Work**
   - Clarify explicitly supported project values and designer stance
   - Situate the work within a social context
2. **Empirical Investigation**
   - Identify key direct and indirect stakeholders
   - Elicit potential values from stakeholders
   - Systematically identify benefits and harms for stakeholders
   - Refine the social context
3. **Conceptual Investigation**
   - Develop working definitions of key values and identify potential value tensions
   - Define technical and technological success objectives
   - Map tensions to success objectives
4. **Technical Investigation**
   - Identify choice points where the design team has the mandate, control, or power to intervene
   - Build technological and social interventions
5. **Monitor and Respond to Change Over Time**

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.

VSD IS NOT AN ALGORITHM

We've provided example VSD questions and steps to help clarify your thinking

But, fundamentally VSD is an outlook and a process

- VSD is not an algorithm
- There is no design recipe for VSD
- There is no way to *#include vsd.h* or *import VSD*

Committing to VSD means being thoughtful and agile

- No single right answer to complex ethical and moral questions…
- But there are lots of wrong answers

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

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

GO FORTH AND BUILD RESPONSIBLY

Questions? Contact Professor Christo Wilson (c.wilson@northeastern.edu)

Khoury College of Computer Sciences, Northeastern University

SOURCES

Value Sensitive Design: Shaping Technology with Moral Imagination (The MIT Press). Batya Friedman and David G. Hendry, 2019. https://www.amazon.com/Value-Sensitive-Design-Technology-Imagination/dp/0262039532/

The algorithms that detect hate speech online are biased against black people (Recode). Shirin Ghaffary, 2019. https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter

YouTube Said It Was Getting Serious About Hate Speech. Why Is It Still Full of Extremists? (Gizmodo). Aaron Sankin, 2019. https://gizmodo.com/youtube-said-it-was-getting-serious-about-hate-speech-1836596239