IEEE WIE ILC
2021

Embrace Ethics To Make Trustworthy Tech

Carol J. Smith
Sr. Research Scientist - Human-Machine Interaction, CMU's SEI
Adjunct Instructor, CMU's Human-Computer Interaction Institute

Twitter: **@carologic**

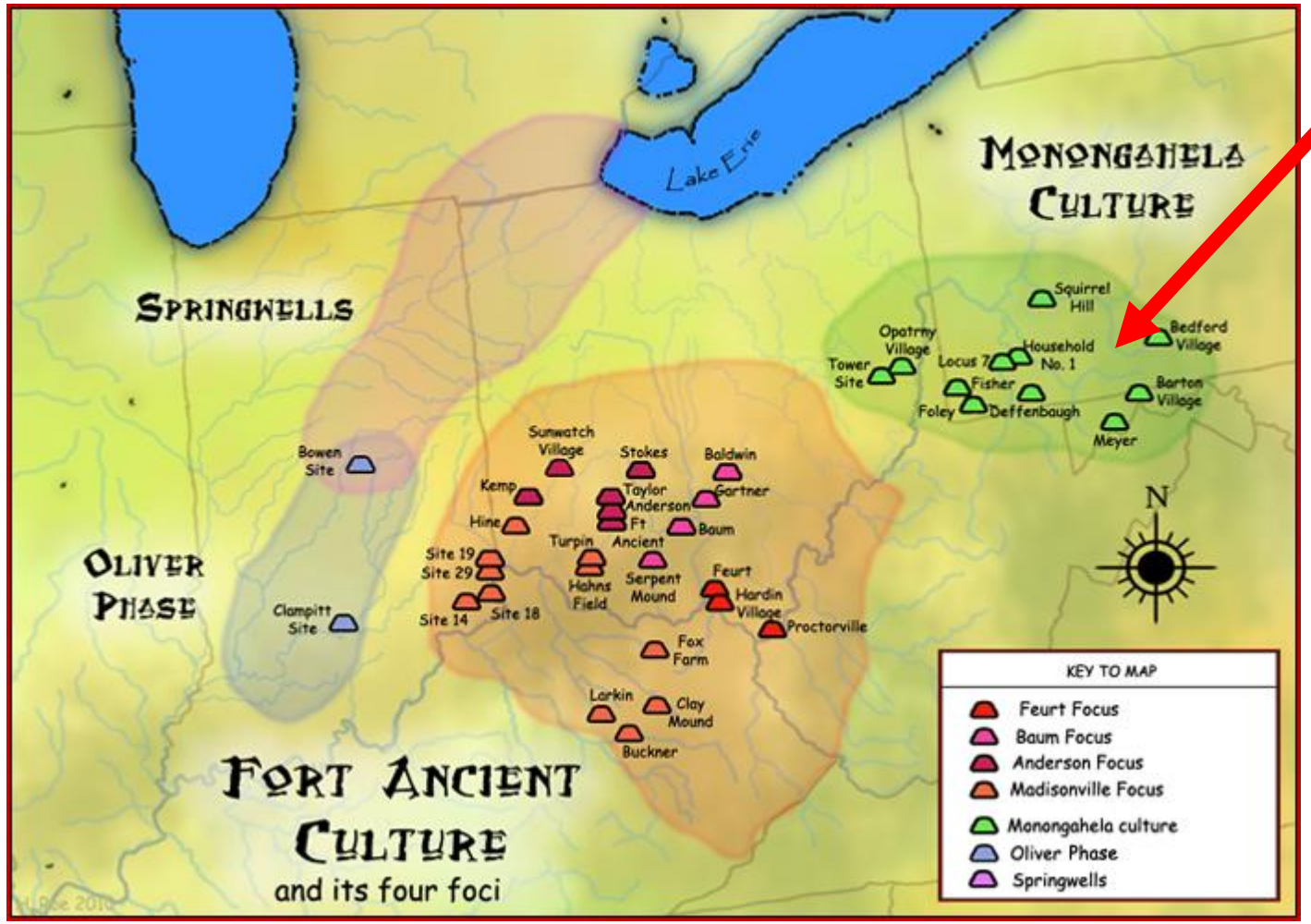
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Copyright Statement

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

DM21-0322

Acknowledgement: The Land I Speak On



Map by Herb Roe via Wikipedia https://en.wikipedia.org/wiki/Monongahela_culture

Land of Monongahela, Adena and Hopewell Nations;
Seneca, Lenape and Shawnee lands;
Osage, Delaware and Iroquois lands.

Now known as Pittsburgh, PA, USA.

AI and Emerging Technology

Great potential - develop with caution

abc NEWS

Ring security camera hacks see homeowners subjected to racial abuse, ransom demands

A spate of incidents has seen homeowners in four states fall victim to hackers.

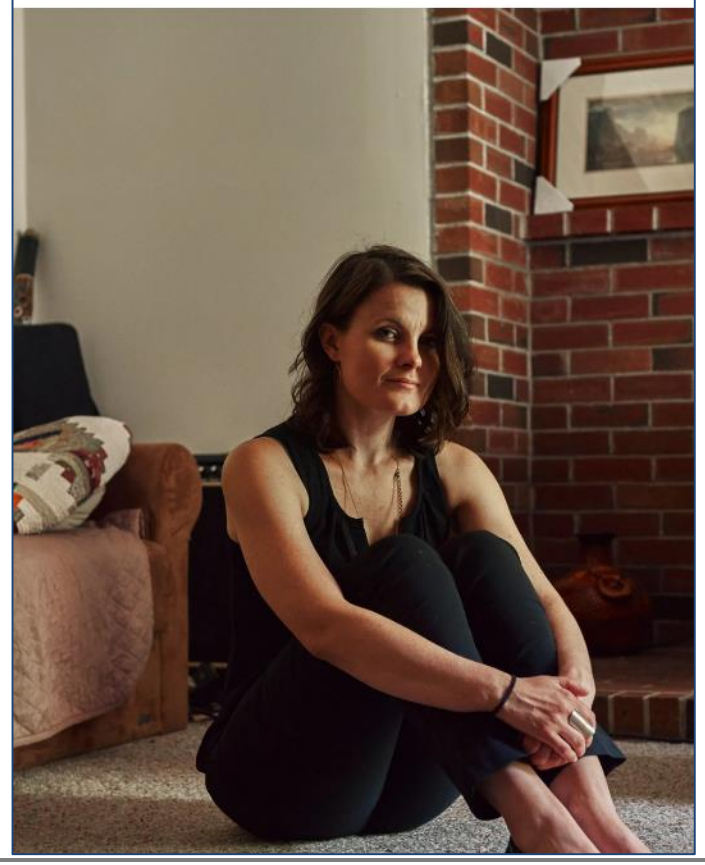
By Mark Hanrahan
December 12, 2019, 9:56 PM • 7 min read

Ring camera systems being hacked

Multiple U.S. families have reported incidents of Ring camera systems being hacked in recent days.

The New York Times

Thermostats, Locks and Lights: Digital Tools of Domestic Abuse



FAST COMPANY

07-07-17

Nest Founder: "I Wake Up In Cold Sweats Thinking, What Did We Bring To The World?"

Tony Fadell, one of the minds behind the iPod and the iPhone, mulls design's unintended consequences.

[Photos: Constantin Renner/EyeEm/Getty Images, davide ragusa/Unsplash]

Assuming That Math Reduces Bias

Vetting applicant resumes

REUTERS

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / 10 MONTHS AGO

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.

Embedding Existing Bad Behaviors

Determining interest rates for mortgage lenders

Berkeley News

Research People Campus & community

BUSINESS & ECONOMICS, RESEARCH

Mortgage algorithms perpetuate racial bias in lending, study finds

By Public Affairs, UC Berkeley | NOVEMBER 13, 2018

Tweet Share 150 Email Print

“Full-Self Driving”

Tesla Autopilot in Heavy LA Traffic by Scott Kubo https://youtu.be/m3-QzTFxoUg?t=14

Joy Buolamwini, Algorithmic Justice League

Coded gaze

“Data is a function of our history…
The past dwells within our algorithms…
Showing us the inequalities that have always been there.”

Photo: Joy Buolamwini on The Open Mind: Algorithmic Justice. Jan 12, 2019. https://www.youtube.com/watch?v=hwHnXdoSSFY

To be biased, is to be human

Bias are shortcuts, to avoid risk and simplify problems.

Not inherently bad,
may be misapplied

Implicit = invisible

Not necessarily in sync
with our conscious beliefs

Can be managed and changed

Talk about biases in non-threatening, productive ways

Biased due to…

Social class

Resource availability

Education

Race, gender, sexuality

Culture, theology, tradition

More…

All systems have some form of bias

Complete objectivity is misleading.

Bias can have purpose and can be helpful.

The goal is to reduce unintended and/or harmful bias.

Diverse, talented and multi-disciplinary

Bringing their varied skills sets, problem framing approaches, and knowledge together.

- Gender, race, culture
- Education (school, program, etc.)
- Experiences
- Thinking process, skill set
- Age, disability and health status, and more…

Representatively diverse leadership for retention

Photo by Christina @ wocintechchat.com on Unsplash
https://unsplash.com/@wocintechchat?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText

Not lowering bar

——————— extending

Great Minds Think Different

High value in diverse teams

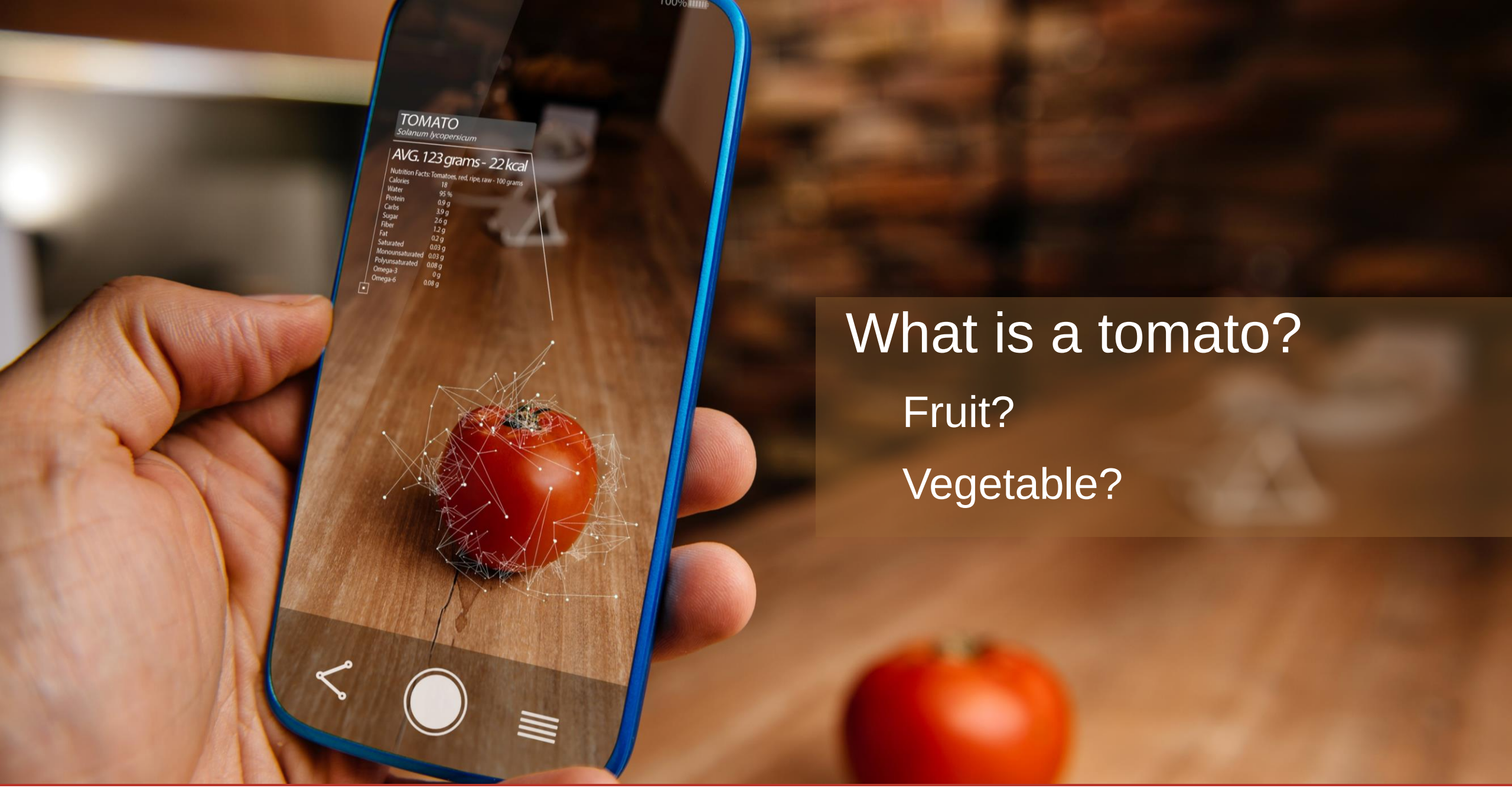
Focus more on facts

Process facts more carefully

More innovative

"…become more aware of their own potential biases"

Why Diverse Teams Are Smarter. Harvard Business Review. https://hbr.org/2016/11/why-diverse-teams-are-smarter

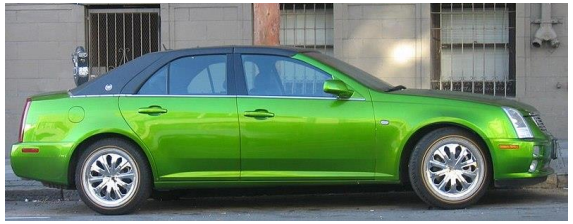


What is a tomato?

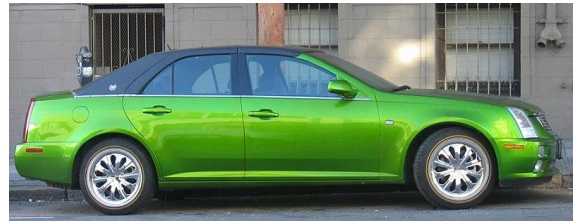
- Fruit?
- Vegetable?

Bias in Emerging Technology – Image Recognition

Training data

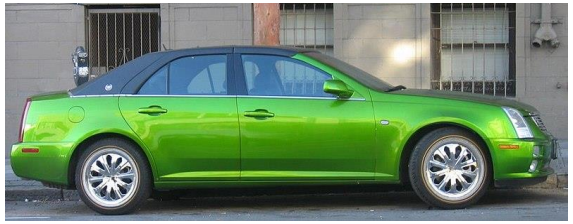


Data encountered



Only know what taught

Training data



Unrepresentative
or incomplete training data

Data encountered

Unlikely to recognize

Responsible, Intentional Design

Just because you can, doesn't mean you should.

Adopt Technology Ethics

- Harmonize cultural variations
- Balance to pace of change, industry pressure
- Explicit permission to consider and question breadth of implications

uxpa international

acm Association for Computing Machinery

AINOW INSTITUTE

Microsoft

Google

< >
Montréal Declaration
Responsible AI_
</ >
An initiative of Université de Montréal

U.S. DEPT OF DEFENSE

ETHICAL OS

Coalesce on Shared Set of Technology Ethics

1. Well-being
2. Respect for autonomy
3. Protection of privacy and intimacy
4. Solidarity
5. Democratic participation
6. Equity
7. Diversity inclusion
8. Prudence
9. Responsibility
10. Sustainable development

Montréal Declaration for a responsible development of artificial intelligence.
https://www.montrealdeclaration-responsibleai.com/the-declaration

Diverse, inclusive leaders

Diverse, Multi-Disciplinary Teams

Shared Tech Ethics

UX Framework

Making Trustworthy Technology

Trust is earned.
Trust must be appropriate.

Early, purposeful work

What is the challenge being face?

For whom? What are their needs?

What kind of improvements are expected?

What might a machine do better or faster?

What is not going to be improved (out of scope)?

How do we get there?

Trustable,
Ethical
Technology

Montréal Declaration for a responsible development of artificial intelligence.
https://www.montrealdeclaration-responsibleai.com/the-declaration

Conversations for Understanding

Difficult Topics

- What do we value?
- Who could be hurt?
- What lines won't our AI cross?
- How are we shifting power?*
- How will we track our progress?

*"Don't ask if artificial intelligence is good or fair, ask how it shifts power." Pratyusha Kalluri. https://www.nature.com/articles/d41586-020-02003-2

Photo by Pam Sharpe https://unsplash.com/@msgrace?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText On Unsplash - https://unsplash.com/s/photos/business-woman-smiling?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText

New uncomfortable work

“*Be uncomfortable*”

- Laura Kalbag

Ethical design is not superficial.

Activate curiosity

UX research methods to activate curiosity:

- Abusability Testing (Dan Brown)
- "Black Mirror" Episodes (Casey Fiesler)
  (inspired by British dystopian sci-fi tv series of same name)
- Flip it to test it
- Implicit Association Test from Harvard University

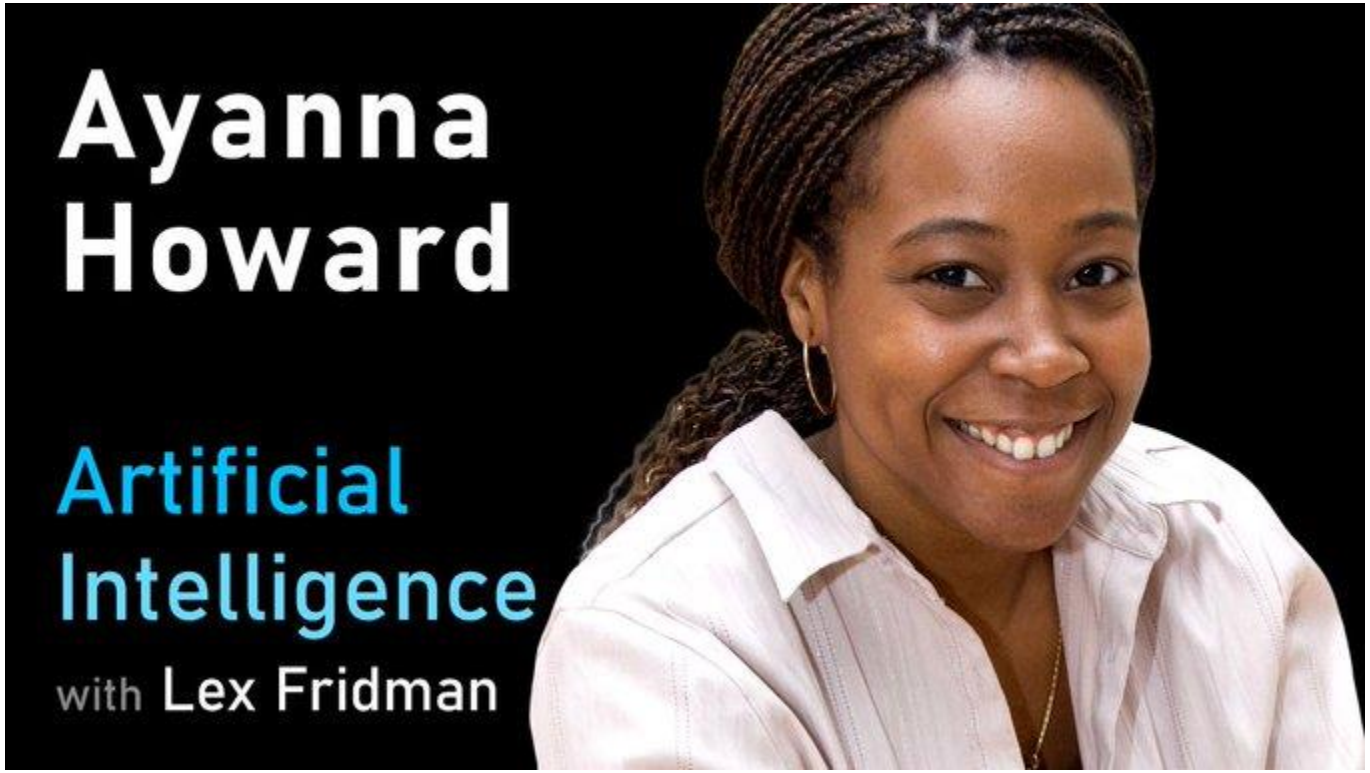
Speculate about system misuse and abuse

- What are potential unintended/unwanted consequences?

More methods to "Outsmart Your Own Biases.": https://hbr.org/2015/05/outsmart-your-own-biases
Implicit Association Test (IAT): https://implicit.harvard.edu/implicit/takeatest.html

Reward team members for finding ethics bugs

Dr. Ayanna Howard
- on the Artificial Intelligence Podcast with Lex Fridman



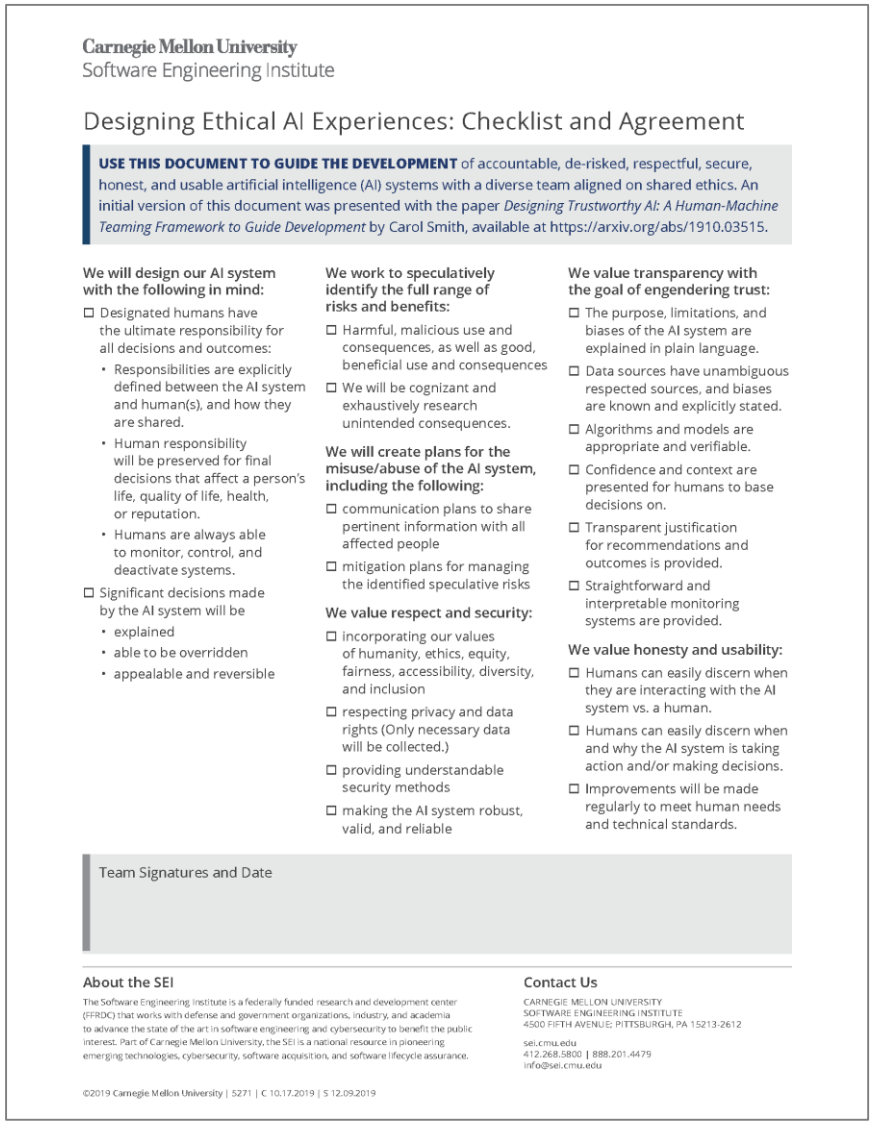
Prompt conversations

Pair Checklist with Technical Ethics

- Bridges gap between "do no harm" and reality

Reduce risk and unwanted bias

Support inspection and mitigation planning



Carnegie Mellon University
Software Engineering Institute

Designing Ethical AI Experiences: Checklist and Agreement

USE THIS DOCUMENT TO GUIDE THE DEVELOPMENT of accountable, de-risked, respectful, secure, honest, and usable artificial intelligence (AI) systems with a diverse team aligned on shared ethics. An initial version of this document was presented with the paper *Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development* by Carol Smith, available at https://arxiv.org/abs/1910.03515.

We will design our AI system with the following in mind:

- ☐ Designated humans have the ultimate responsibility for all decisions and outcomes:
  - Responsibilities are explicitly defined between the AI system and human(s), and how they are shared.
  - Human responsibility will be preserved for final decisions that affect a person's life, quality of life, health, or reputation.
  - Humans are always able to monitor, control, and deactivate systems.
- ☐ Significant decisions made by the AI system will be
  - explained
  - able to be overridden
  - appealable and reversible

We work to speculatively identify the full range of risks and benefits:

- ☐ Harmful, malicious use and consequences, as well as good, beneficial use and consequences
- ☐ We will be cognizant and exhaustively research unintended consequences.

We will create plans for the misuse/abuse of the AI system, including the following:

- ☐ communication plans to share pertinent information with all affected people
- ☐ mitigation plans for managing the identified speculative risks

We value respect and security:

- ☐ incorporating our values of humanity, ethics, equity, fairness, accessibility, diversity, and inclusion
- ☐ respecting privacy and data rights (Only necessary data will be collected.)
- ☐ providing understandable security methods
- ☐ making the AI system robust, valid, and reliable

We value transparency with the goal of engendering trust:

- ☐ The purpose, limitations, and biases of the AI system are explained in plain language.
- ☐ Data sources have unambiguous respected sources, and biases are known and explicitly stated.
- ☐ Algorithms and models are appropriate and verifiable.
- ☐ Confidence and context are presented for humans to base decisions on.
- ☐ Transparent justification for recommendations and outcomes is provided.
- ☐ Straightforward and interpretable monitoring systems are provided.

We value honesty and usability:

- ☐ Humans can easily discern when they are interacting with the AI system vs. a human.
- ☐ Humans can easily discern when and why the AI system is taking action and/or making decisions.
- ☐ Improvements will be made regularly to meet human needs and technical standards.

Team Signatures and Date

About the SEI

The Software Engineering Institute is a federally funded research and development center (FFRDC) that works with defense and government organizations, industry, and academia to advance the state of the art in software engineering and cybersecurity to benefit the public interest. Part of Carnegie Mellon University, the SEI is a national resource in pioneering emerging technologies, cybersecurity, software acquisition, and software lifecycle assurance.

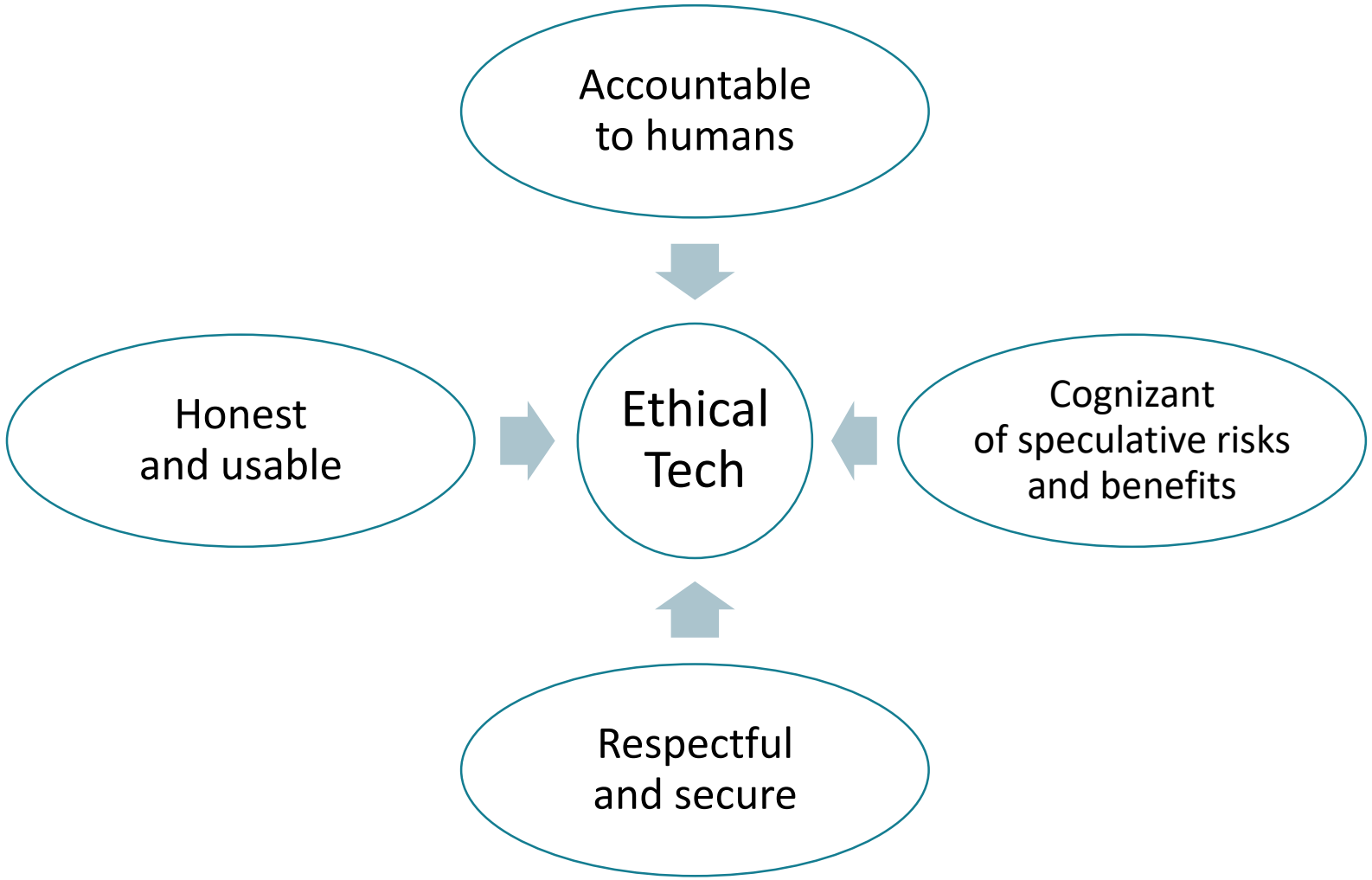
Contact Us

CARNEGIE MELLON UNIVERSITY
SOFTWARE ENGINEERING INSTITUTE
4500 FIFTH AVENUE; PITTSBURGH, PA 15213-2612

sei.cmu.edu
412.268.5800 | 888.201.4479
info@sei.cmu.edu

Checklist and Agreement - Downloadable PDF:
https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620

UX Framework for Designing Trustworthy Tech



RightStaff Scenario

AI shift scheduling system

Users: Store managers of fast-food restaurants

Goals of RightStaff:

- Faster staffing decisions and scheduling
- Reduced bias of shift selection

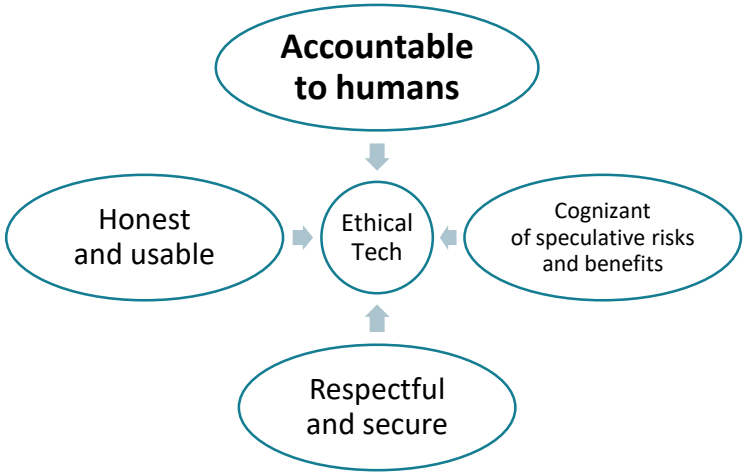
Accountable to Humans

Ensure humans have ultimate control

- Able to monitor and control risk

Human responsibility for final decisions

- Person's life
- Quality of life
- Health
- Reputation



“Ensure humans can unplug the machines”

– Grady Booch

TED Talk, Grady Booch, Scientist, Philosopher, IBM’er

https://www.ted.com/talks/grady_booch_don_t_fear_superintelligence

Significant decisions

Significant decisions made by the AI system will be

- explained
- able to be overridden
- appealable and reversible

RightStaff

- Manager able to reschedule people as needed

Responsibilities explicitly defined

Between AI system and human(s)

RightStaff (AI System or Manager?)

- Picks employees to schedule?
- Defines shifts?
- Method to integrate new information?
  - Sick time
  - Resignations

Abusability Testing

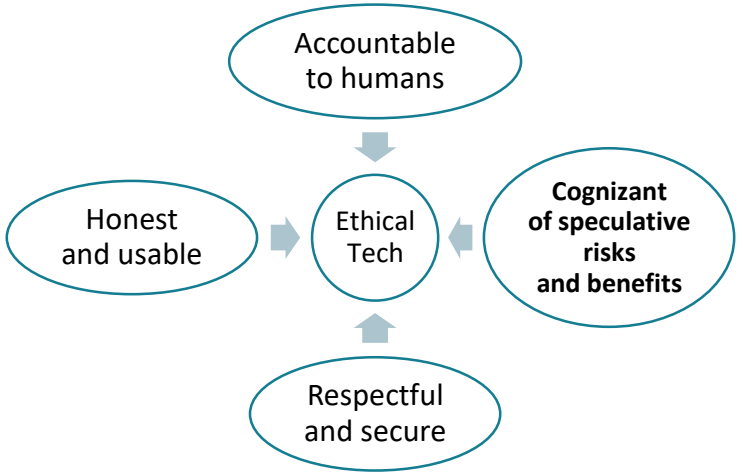
Feature added to enable RightStaff to turn off by itself

- What are limits to functionality?
- How could this be abused/misused?
- Implications?
- Risks?

Cognizant of Speculative Risks and Benefits

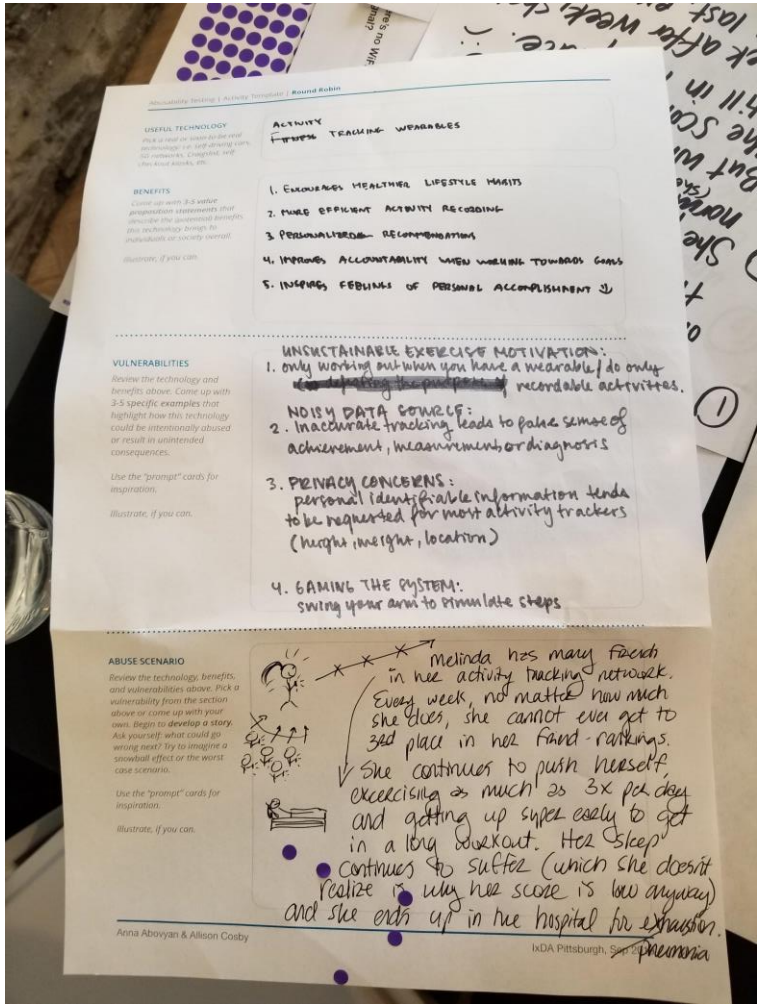
Identify full range of

- Harmful, malicious use, as well as good, beneficial use
- Blind spots and unwanted/unintended consequences



Speculative: Conduct UX research and activate curiosity

- Speculate about misuse and abuse
- Potential severe abuse and consequences
- Perspective of people in frequently marginalized groups
- “Black Mirror” episodes



Bias Scenario

- RightStaff begins prioritizing people with easier schedules
- Managers approve these schedules, reinforcing bias
- People who were previously discriminated against are still discriminated against

Abuse Scenario

- Managers want to avoid providing benefits to employees
- RightStaff adjusted to ensure that no one has full shifts
- No regular staff can keep benefits

- What else?

Speculative: Create communication & mitigation plans

Plan for unwanted consequences

Misuse and abuse of AI system

- Who can report?
- To whom?
- Turn off?
- Who notified?
- Consequences?

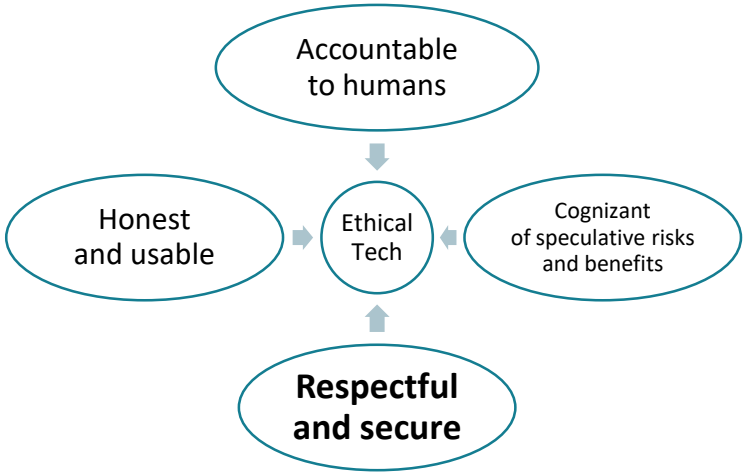
Respectful and Secure

Values of humanity, ethics, equity, fairness, accessibility, diversity and inclusion

Respect privacy and data rights

Make system robust, valid and reliable

Provide understandable security



Respectful and Secure

RightStaff

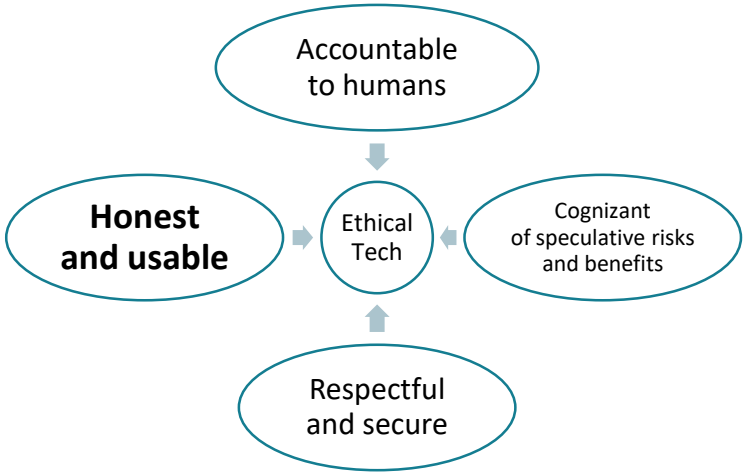
- Who has visibility to reasons for changing schedules?
- How is that information used?
- How is PII* of employees protected?

*PII is Personally Identifiable Information (social security number, address, etc.)

Honest and Usable

Value transparency with the goal of engendering trust

Explicitly state identity as an AI system



Fair: Remove unwanted bias in data

Show awareness of known and desirable bias

Acknowledge issues

Overcommunicate on issues

RightStaff

- System built to reduce the known bias in existing data
- Make it easy to report bias (or prevent it)

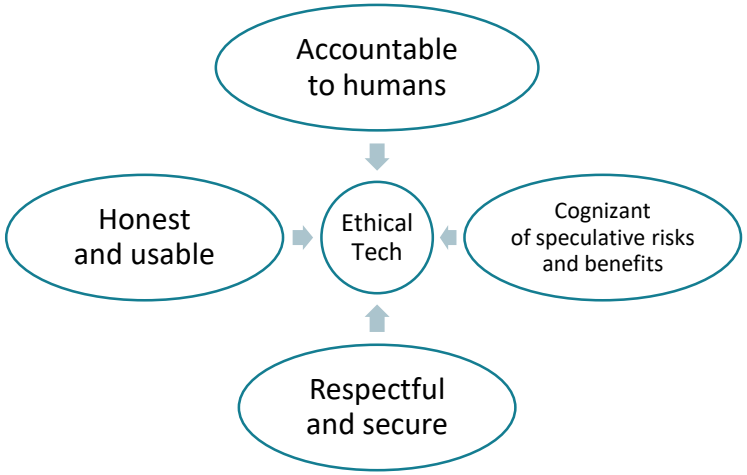
We aren’t perfect, AI won’t be perfect

Empower diverse teams, inclusive environments

Adopt technical ethics

Encourage deep conversations (Checklist)

Activate curiosity; be speculative; imaginative



Evangelize for human values

Ethical. Transparent. Fair.

Carol J. Smith

Twitter: @carologic

LinkedIn: https://www.linkedin.com/in/**caroljsmith**/

Carnegie Mellon University's Software Engineering Institute,
Emerging Technology Center