Carnegie Mellon University
Software Engineering Institute

Trustworthy AI Systems

STEAM BLOOM, AUGUST 2023

Carol J. Smith
Carnegie Mellon University, Software Engineering Institute
Sr. Research Scientist, Human-Machine Interaction

Copyright Statement

DM23-0856

Research to reduce risk and avoid waste

Carnegie Mellon University Software Engineering Institute

Software and AI systems

- Decision support
- Recommender systems
- Large Language Models (LLMs)
- Voice recognition (smart speakers)
- Speech to text/text to speech (live captions)
- And more…

Lack of research results in…

Widely used algorithm for follow-up care in hospitals is racially biased, study finds

PBS News Hour Weekend

Amazon scraps secret AI recruiting tool that showed bias against women

Reuters

The IRS directed 7 million Americans to sign up with ID.me face-scan service, according to congressional letter

The Washington Post

Berkeley News

Mortgage algorithms perpetuate racial bias in lending, study finds

The New York Times

Thermostats, Locks and Lights: Digital Tools of Domestic Abuse

THE WALL STREET JOURNAL.

Slack Backtracks on New Way to Message Over Harassment Concern

Carnegie Mellon University Software Engineering Institute

"The biggest waste of all is building something no one wants"

Eric Ries @ericries via @MelBugai on Twitter at LeanStartupMI in 2011

Everything is Designed

Carnegie Mellon University
Software Engineering Institute

Adherence to Consistency

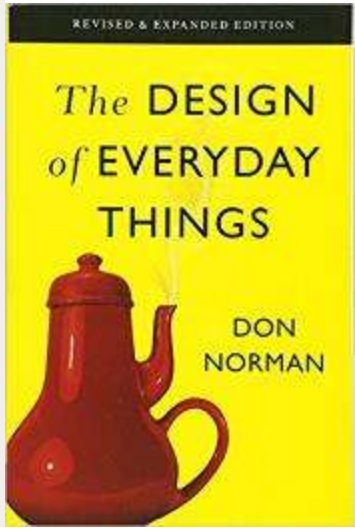
https://www.reddit.com/r/dangerousdesign/comments/3s75gz/cooking_spray_insect_killer/

Change over time

https://boingboing.net/2014/02/25/the-story-behind-the-too-coo.html

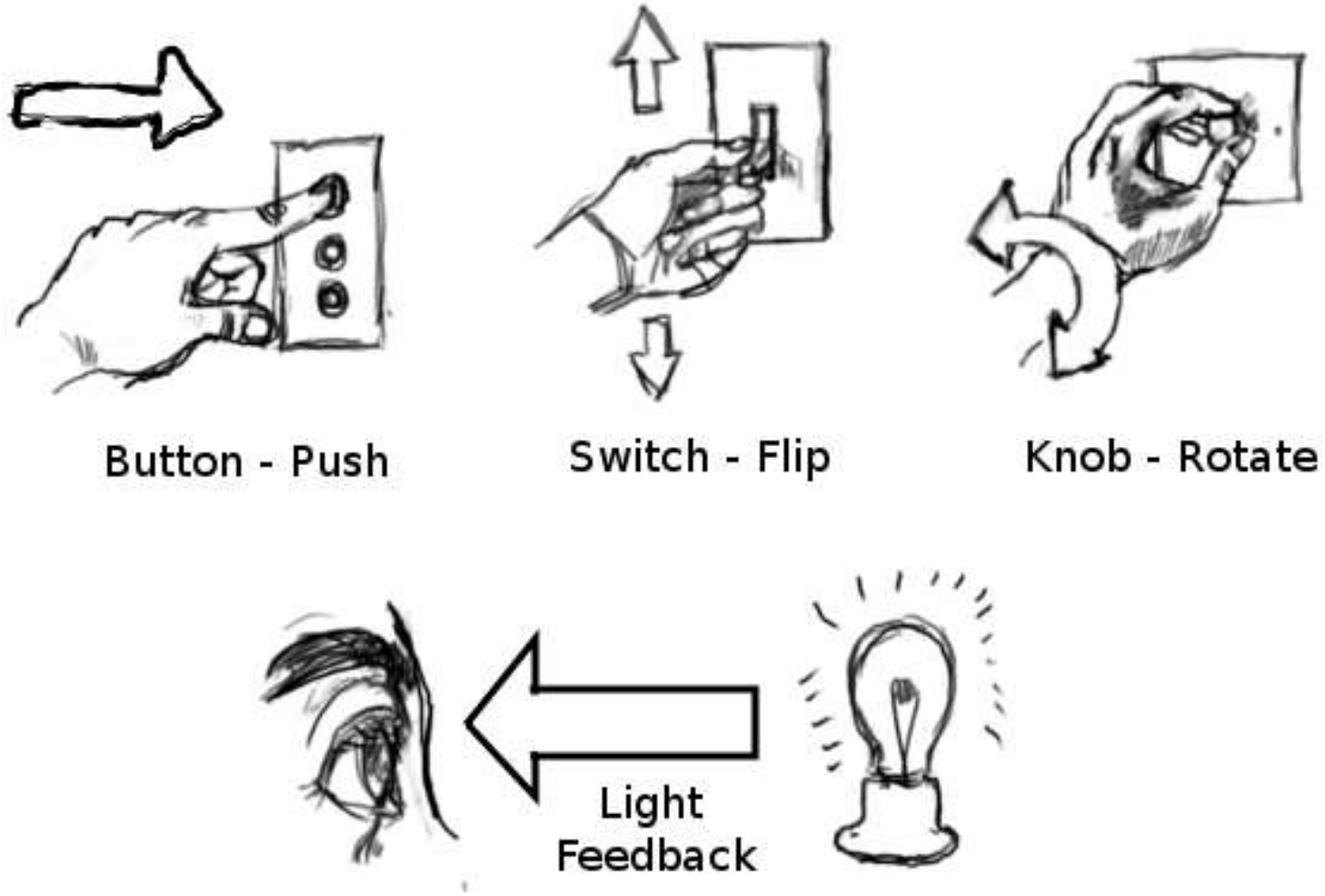
Long tradition – designing better experiences

- Human-Computer Interaction
- User Experience
- Usability
- Interaction Design



The "masochist's teapot", as borrowed by Don Norman for his book *The Design of Everyday Things*. Image: http://playerside.blogspot.com/2012/12/the-design-of-everyday-gaming-feedback.html

Make informed design choices



Drawings of Affordance: http://paaralan.blogspot.com/2010/09/affordance-and-educational-games.html

Carol J. Smith

Carnegie Mellon University Software Engineering Institute

Software Engineering Institute

AI Division Staff

- Sr. Research Scientist, human-machine interaction
- AI/ML, autonomy, emerging technologies
- Government agencies

Adjunct Instructor

- Interaction Design Overview
- Human-centered design
- Prototyping
- Design and iteration

My Work

Carnegie Mellon University Software Engineering Institute

Focus on interesting and influential topics

- Someone has an idea
- Explore options
- Plan a research project

Projects are long (1+ years)

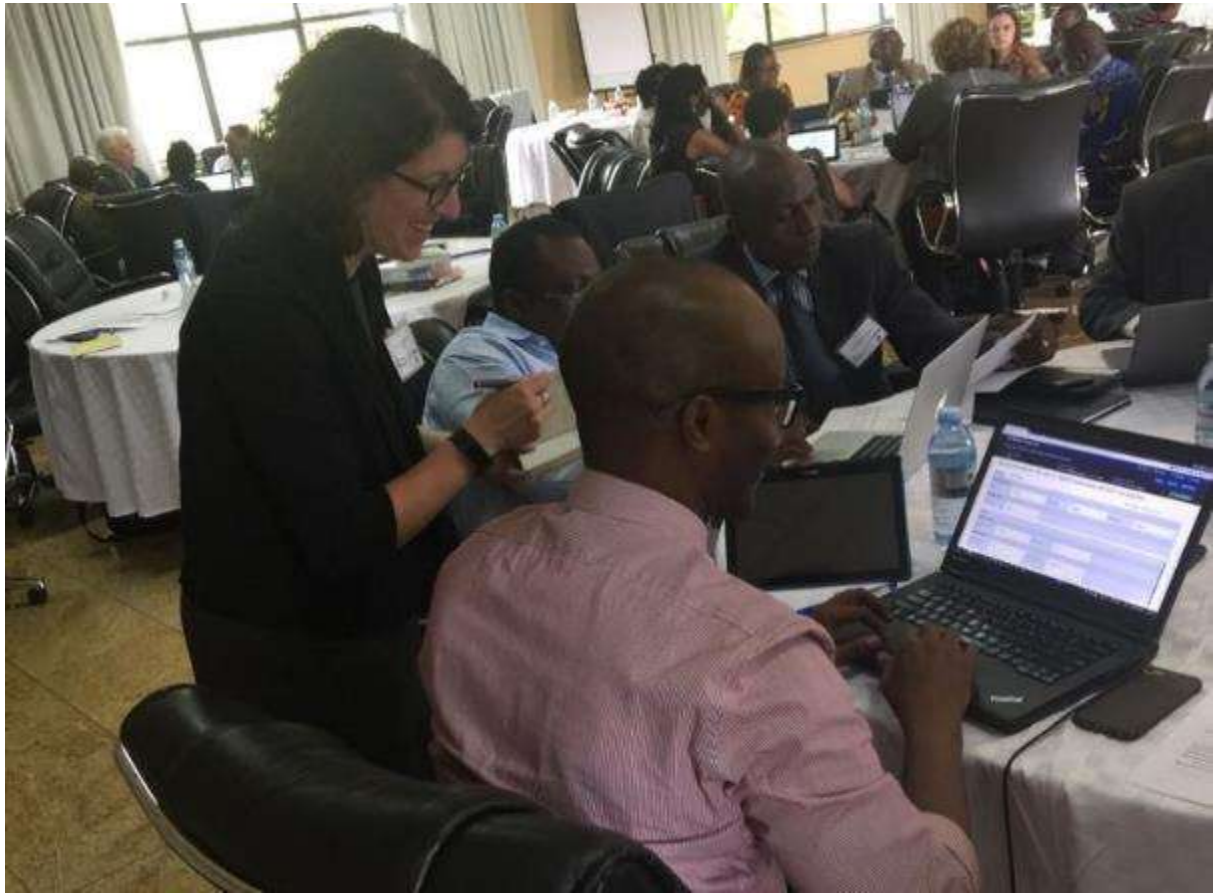
Work with software engineers, machine learning researchers, scientists, and more…

Explore the situation

- Who would use the system?
- What would they do?
- What are their needs?

- Conduct Research Observations, Interviews

Observe use of system and environment

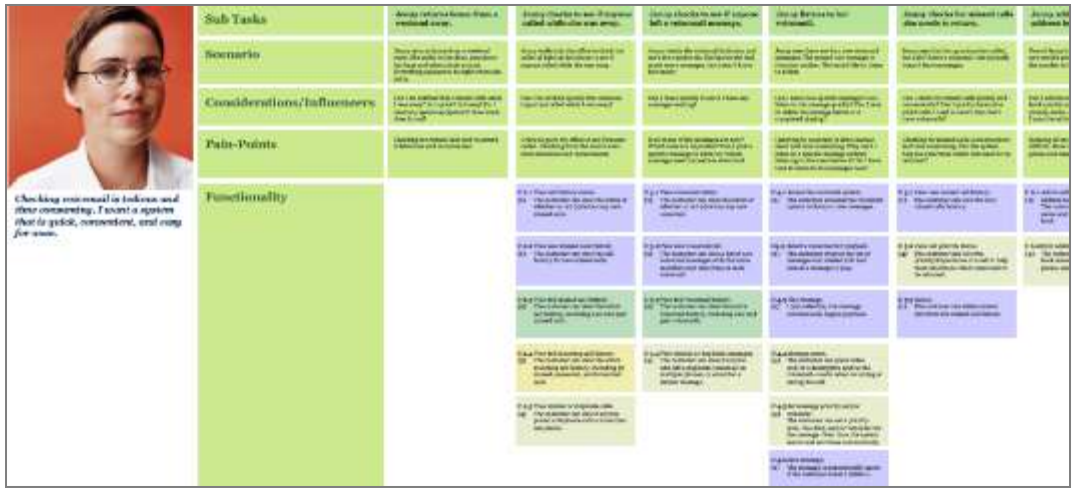


Analyze Information

Carnegie Mellon University Software Engineering Institute

Map out information
- Look for patterns
- What people do
- Why they do it?

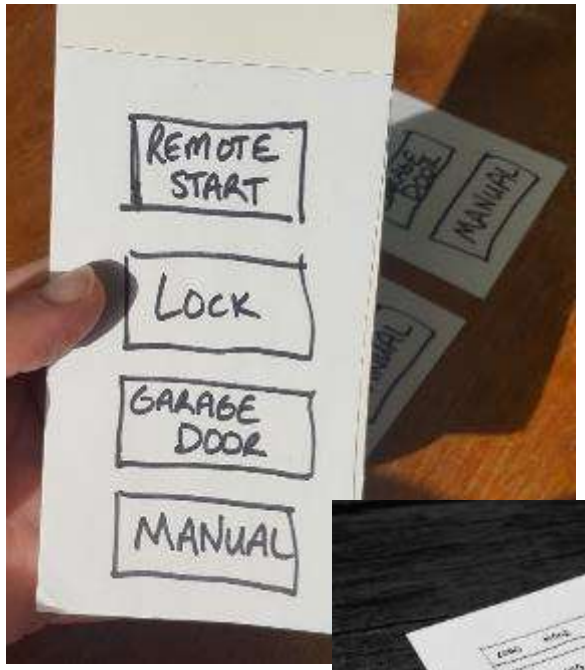
- Create helpful references
- Not “books” for shelving

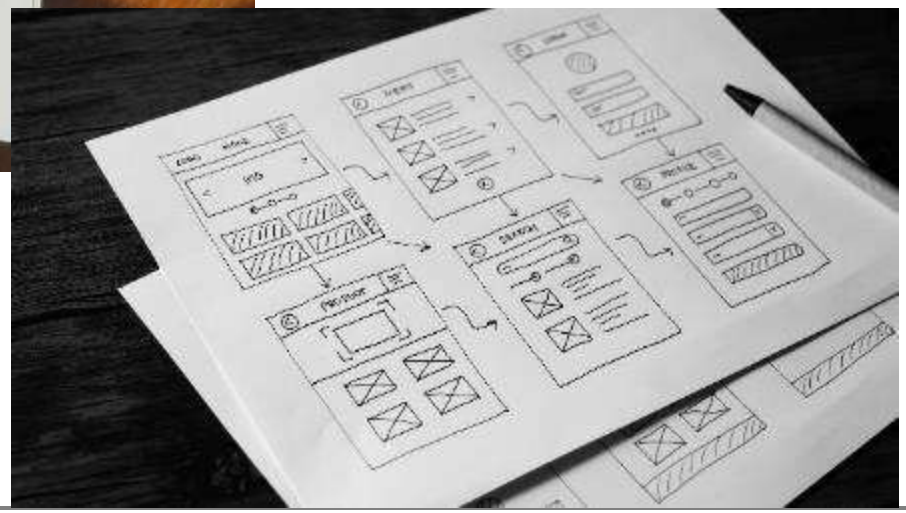


Example of a Task Analysis by Todd Zaki Warfel from his Agile2010 presentation "Opening the Kimono a look behind the design process."

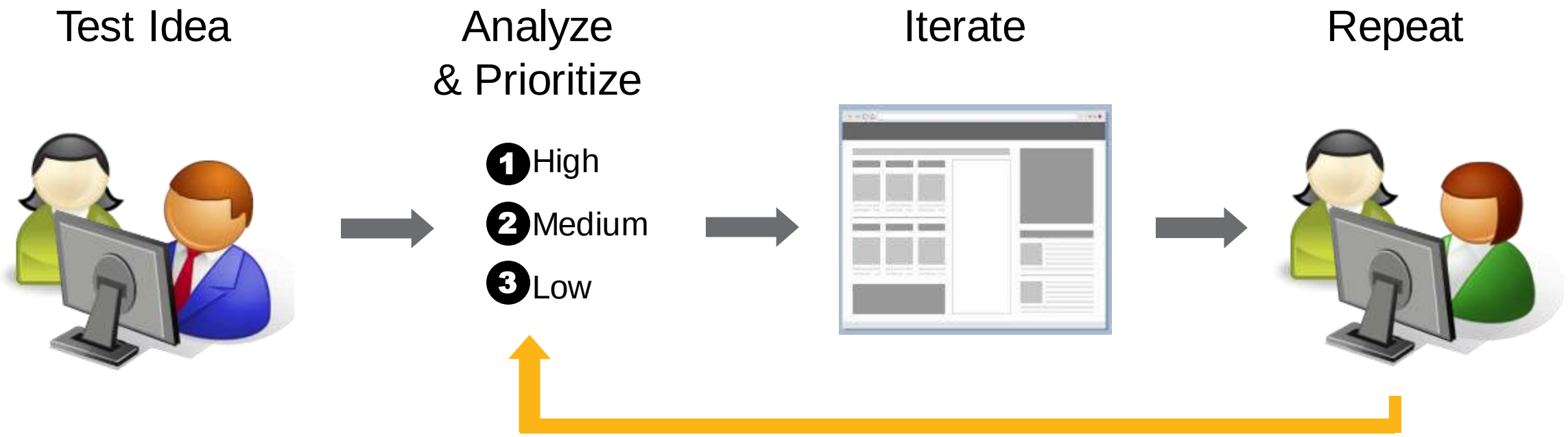
Prototype – Experiments to Learn

- How can I best support the user?
- How do I solve this problem?
- What interaction will meet the end users' needs?





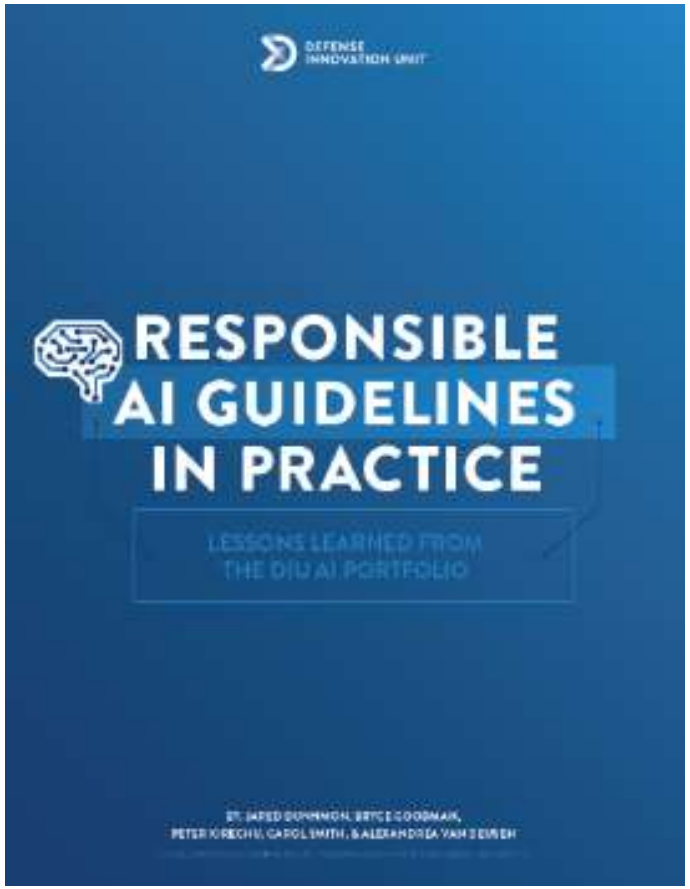
Iterative Cycles: Feedback and Improvement

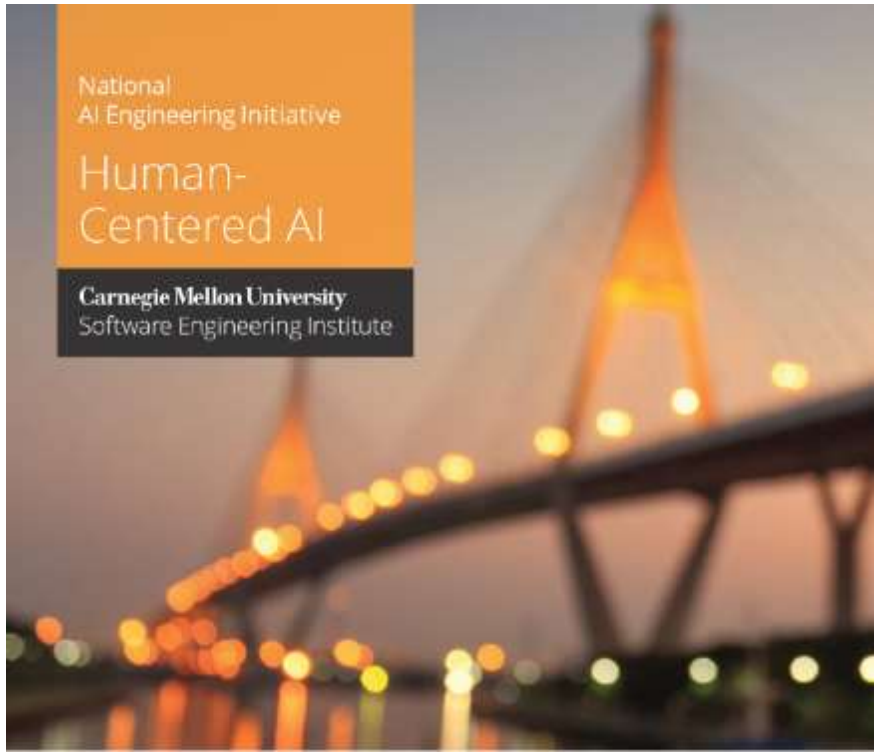


Presentations and Publications


Carnegie Mellon University
Software Engineering Institute

Designing Ethical AI Experiences: Checklist and Agreement



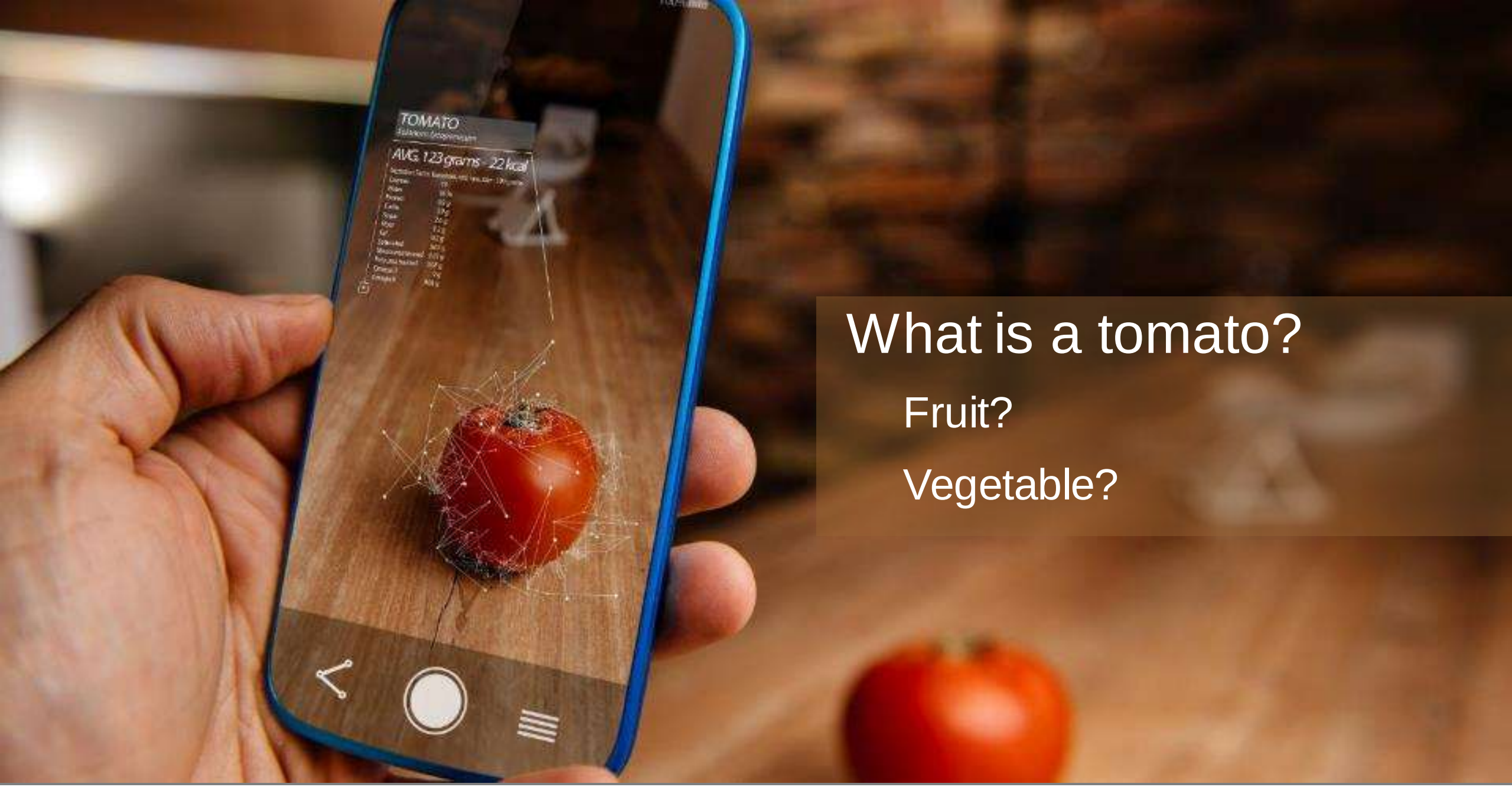


Human-Centered AI, White Paper. June 2021. CMU's Software Engineering Institute. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362

Carol J. Smith. Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development. https://arxiv.org/abs/1910.0351
Checklist and Agreement - Downloadable PDF: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620
Defense Innovation Unit. Artificial Intelligence Portfolio, Responsible AI Guidelines. https://www.diu.mil/responsible-ai-guidelines

Making Responsible Artificial Intelligence

Carnegie Mellon University
Software Engineering Institute



What is a tomato?

Fruit?

Vegetable?

Computer Vision - Image Recognition

Carnegie Mellon University Software Engineering Institute

Train set

Data encountered

Use case courtesy of Dr. Eric Heim, CMU Software Engineering Institute
https://resources.sei.cmu.edu/library/author.cfm?authorid=542374

Only know what taught

Carnegie Mellon University Software Engineering Institute

Train set

Unrepresentative
or incomplete training data

Data encountered

Unlikely to recognize

Carnegie Mellon University
Software Engineering Institute

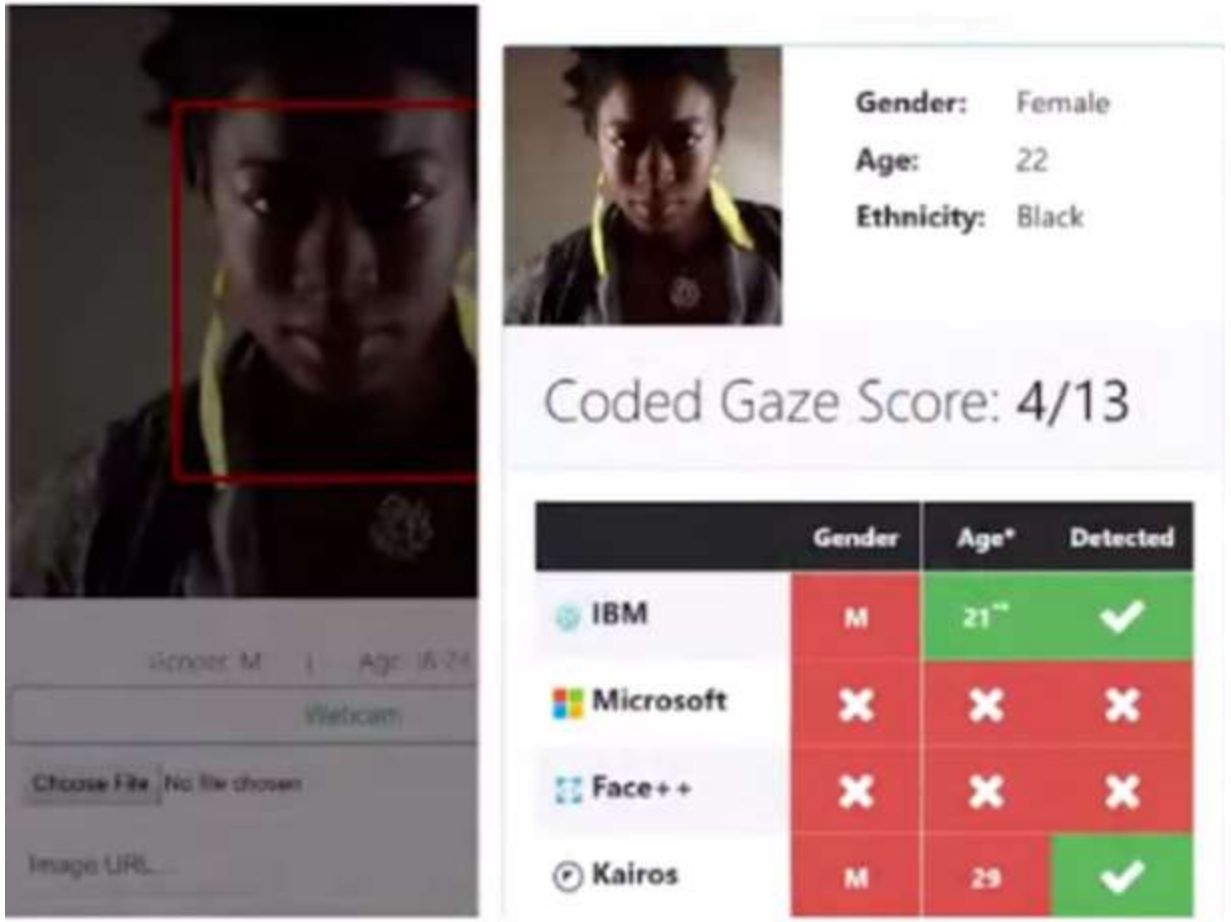
"Data is a function of our history…
The past dwells within…
Showing us the inequalities
that have always been there."

Joy Buolamwini, Algorithmic Justice League
Coded Gaze
Movie: Coded Bias on Netflix

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

Data is biased

- Not inherently neutral.
- Reflect priorities, preferences, and prejudices of people making them



Gender Shades Project by Joy Buolamwini. MIT Media Lab. https://www.media.mit.edu/projects/gender-shades/overview/

24

Algorithms of Oppression, Dr. Safiya Noble

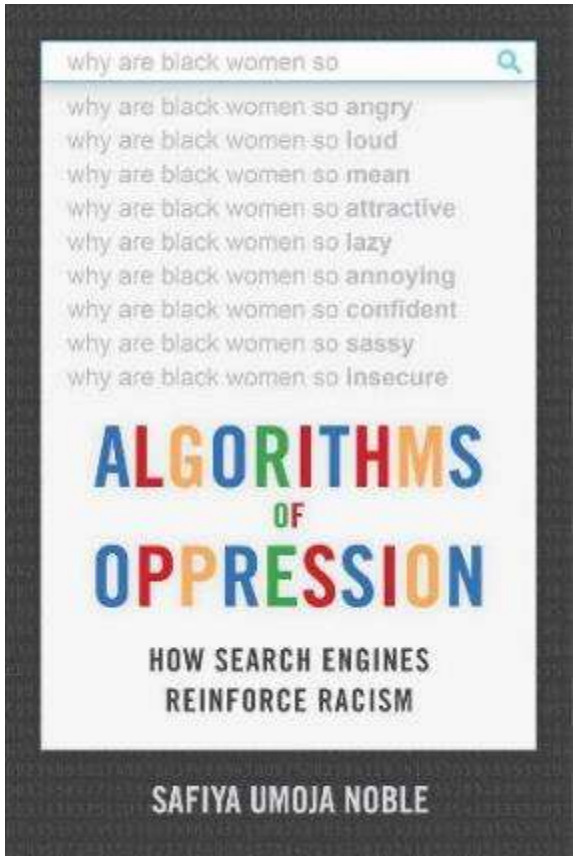


Photo from article: Google This: Algorithmic Oppression. ACLU News & Commentary. February 19, 2021. https://www.aclu.org/news/privacy-technology/google-this-algorithmic-oppression/

25

Carnegie Mellon University Software Engineering Institute

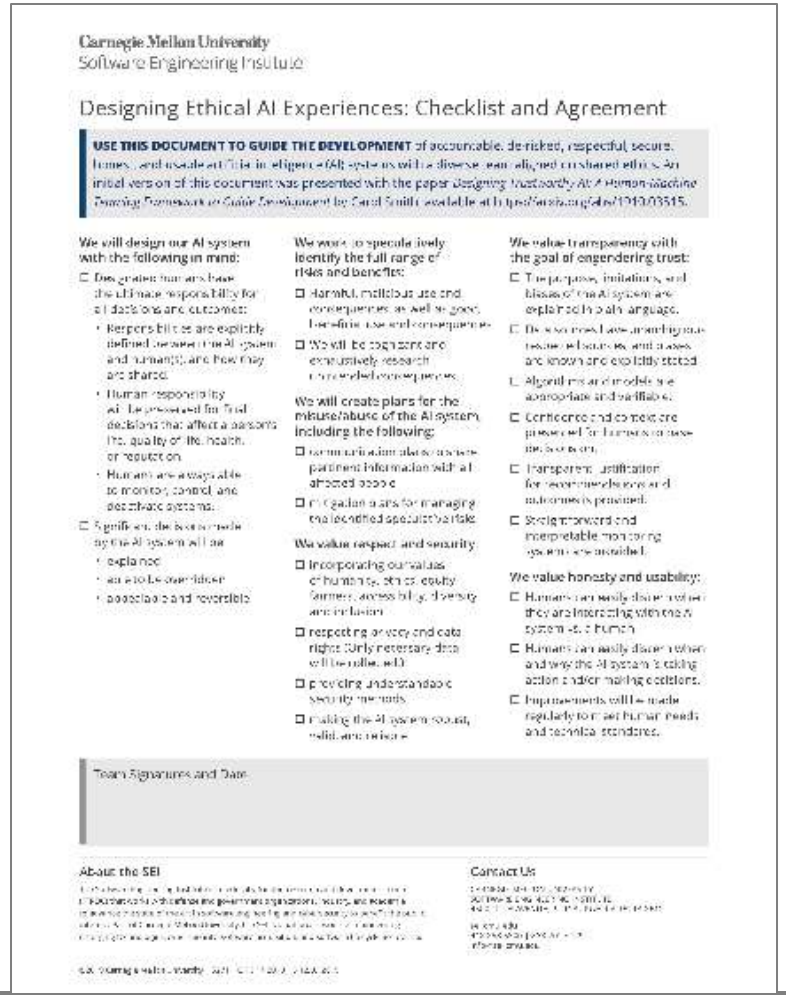
UX Framework to guide AI teams

1) Accountable to humans

2) Cognizant of speculative risks and benefits

3) Respectful and secure

4) Honest and usable



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, human-centered and usable artificial intelligence (AI) systems with a diverse, team-aligned, co-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:

☐ Designers that are held the ultimate responsibility for all decisions and outcomes:
- Responsibilities are explicitly defined (between the AI system and humans), and how they are shared.
- Human responsibility will be presented for final decisions that affect a person's (hs. 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:
- explainable
- able to be overridden
- appealable and reversible

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

☐ Harmful, malicious use and consequences as well as good. Benefits, 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:

☐ [illegible] 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 controls

☐ making the AI system robust, reliable, and resilient

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, [illegible] resources, and biases are known and explicitly stated.

☐ Algorithms are explainable, appropriate, and verifiable.

☐ Confidence and context are presented for [illegible]

☐ Transparent justification for recommendations and outcomes is provided.

☐ Strategy for [illegible] monitoring system use is 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

Contact Us

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?*

*"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 *"How is this ML model shifting power?" @riakall #NeurIPS2019

Photo by Pam On Unsplash

- “Ensure humans can unplug the machines”
- – Grady Booch

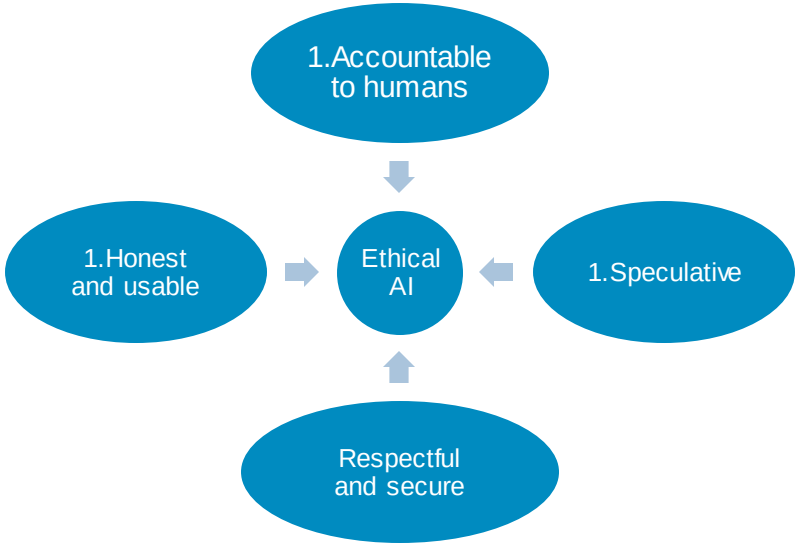
TED Talk, Grady Booch, Scientist, Philosopher, IBM’er
https://www.ted.com/talks/grady_booch_don_t_fear_superintelligence

Carnegie Mellon University Software Engineering Institute

Reward team members for finding ethics bugs

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



Advice

Carnegie Mellon University
Software Engineering Institute

Notice your priorities

- What you can put up with. What you cannot.
- Invest in experiences. In learning.
  - Lots of working days ahead
  - Learn what you can, for as long as you can.
  - Look for fun, interesting work

Not learning? Not being challenged?

- Look/ask for new opportunities.
- Also, change is ok.
- Go and try something new.
- Prepare to tell your story.

Build a manager Voltron

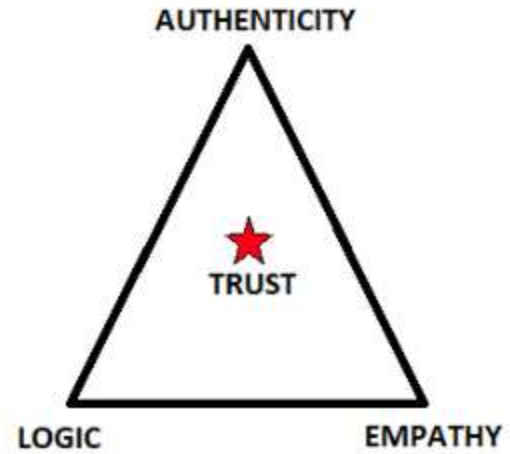
Trust your instincts about people.

Develop diverse manager crew/coaching network.
Look for people who:

- push you out of your comfort zone
- have different levels of experience than you
- are from a different background
- are good at things you're terrible at

When your manager isn't supporting you, build a Voltron by Lara Hogan Originally posted Jan 4, 2018
https://larahogan.me/blog/manager-voltron/

Authenticity. Empathy. Logic.



Understand where you [and others] wobble
- Frances Frei, Harvard Business School professor

Frances Frei, Harvard Business School, https://blog.ted.com/how-to-rebuild-trust-frances-frei-speaks-at-ted2018/

Your Responsibility

- Ask for feedback
- Question status quo
- Stand for ethics – do the right thing
- Talk about bias
- Be inclusive and kind
- Don’t plan too far out

Carnegie Mellon University
Software Engineering Institute

Carol J. Smith

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