

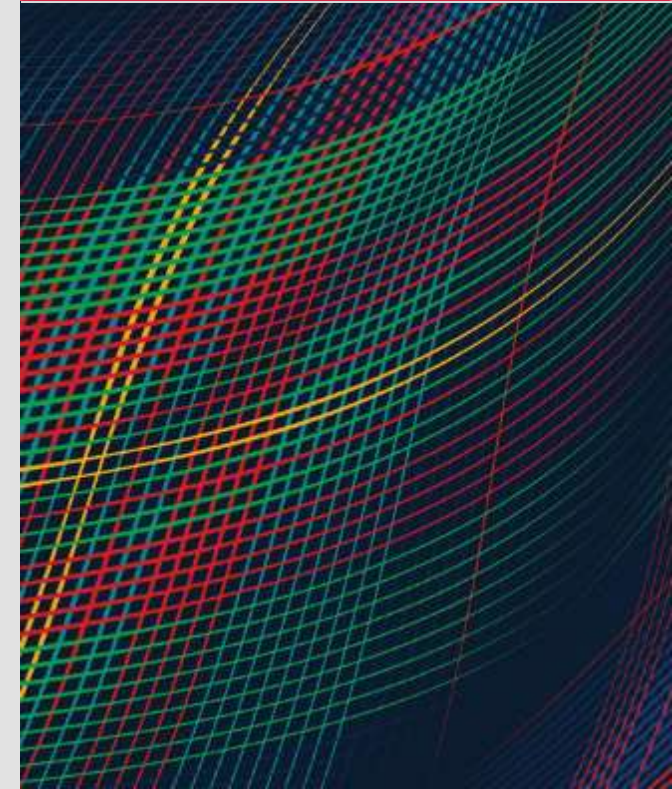
# Trustworthy AI Systems

**STEAM BLOOM, AUGUST 2023**

Carol J. Smith

Carnegie Mellon University, Software Engineering Institute

Sr. Research Scientist, Human-Machine Interaction



# Copyright Statement

Copyright 2023 Carnegie Mellon University.

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.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at [permission@sei.cmu.edu](mailto:permission@sei.cmu.edu).

Carnegie Mellon® is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM23-0856

# Research to reduce risk and avoid waste

## 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

The New York Times

## Thermostats, Locks and Lights: Digital Tools of Domestic Abuse



## Berkeley News Mortgage algorithms perpetuate racial bias in lending, study finds

## THE WALL STREET JOURNAL.

## Slack Backtracks on New Way to Message Over Harassment Concern

**“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

# Adherence to Consistency



[https://www.reddit.com/r/dangerousdesign/comments/3s75gz/cooking\\_spray\\_insect\\_killer/](https://www.reddit.com/r/dangerousdesign/comments/3s75gz/cooking_spray_insect_killer/)

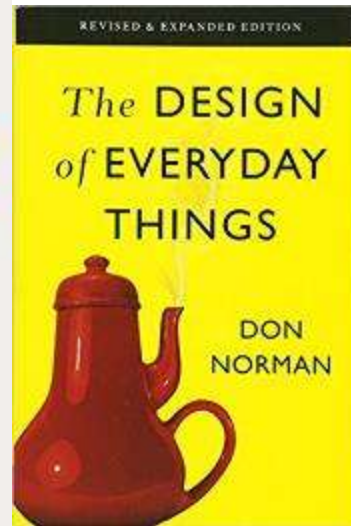
# Change over time



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

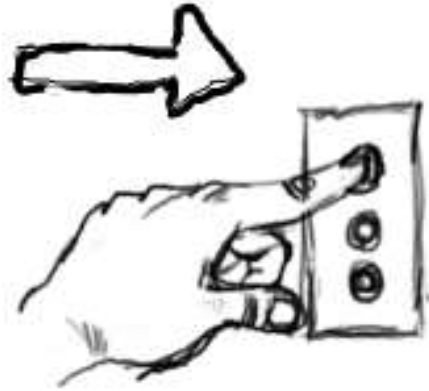
# Long tradition – designing better experiences

- 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



Button - Push



Switch - Flip



Knob - Rotate



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

# Carol J. Smith

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

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

## Map out information

- Look for patterns
- What people do
- Why they do it?
- Create helpful references
- Not “books” for shelving

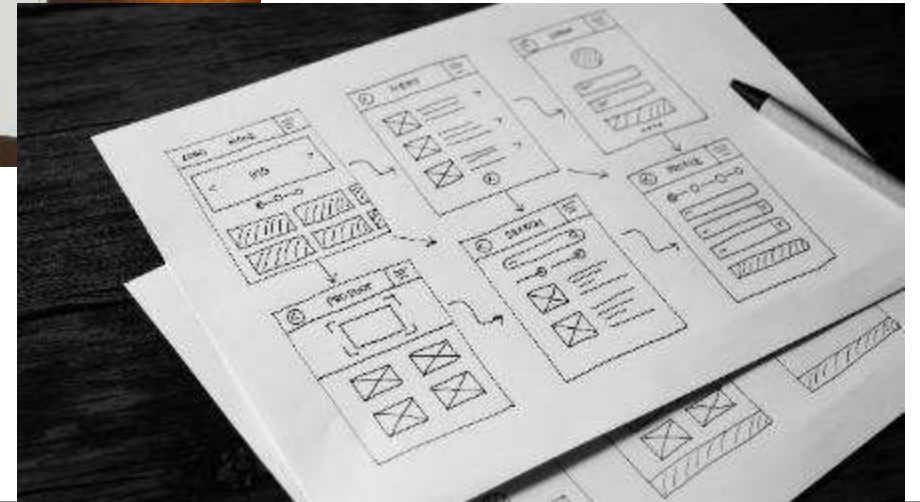
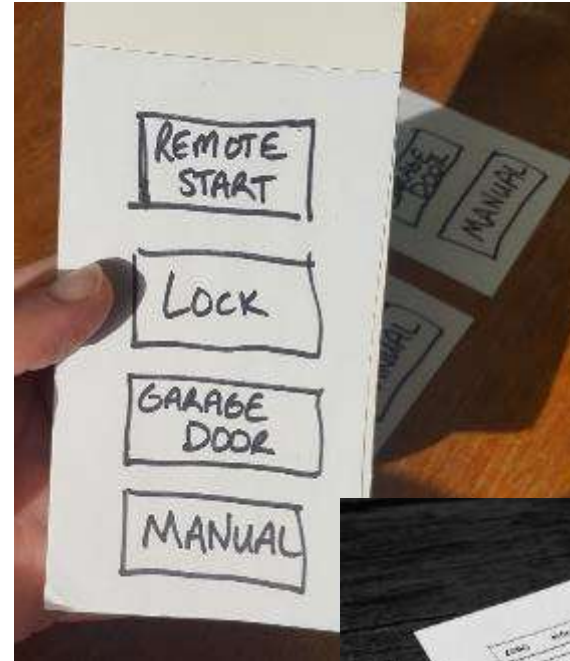
 <p>Checking external factors and their consequences. I want a system that is quick, convenient, and easy for users.</p>	Sub Tasks	Verify that the system has a consistent design.	Verify that the system is easy to use and intuitive.	Verify that the system is secure and reliable.	Verify that the system is scalable and flexible.	Verify that the system is maintainable and upgradeable.	Verify that the system is accessible and usable by all.	Verify that the system is cost-effective and efficient.	
	Scenario	Users can log in to the system and access their account information.	Users can search for products and view product details.	Users can add items to their cart and checkout.	Users can manage their account settings and preferences.	Users can track their orders and receive shipping notifications.	Users can provide feedback and ratings for products.	Users can view their purchase history and receipts.	
	Considerations/Influencers	Users want a simple and intuitive interface.	Users want fast and reliable performance.	Users want a secure and trustworthy system.	Users want a system that can grow with their needs.	Users want a system that is easy to integrate with other tools.	Users want a system that is easy to learn and use.	Users want a system that is easy to support and maintain.	Users want a system that is easy to upgrade and evolve.
	Pain-Points	Users are frustrated with the current system's complexity.	Users are frustrated with the current system's slow performance.	Users are frustrated with the current system's security concerns.	Users are frustrated with the current system's lack of flexibility.	Users are frustrated with the current system's poor integration with other tools.	Users are frustrated with the current system's steep learning curve.	Users are frustrated with the current system's high support costs.	Users are frustrated with the current system's limited upgrade options.
	Functionality	Users can log in to the system and access their account information.	Users can search for products and view product details.	Users can add items to their cart and checkout.	Users can manage their account settings and preferences.	Users can track their orders and receive shipping notifications.	Users can provide feedback and ratings for products.	Users can view their purchase history and receipts.	Users can view their account balance and payment options.

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

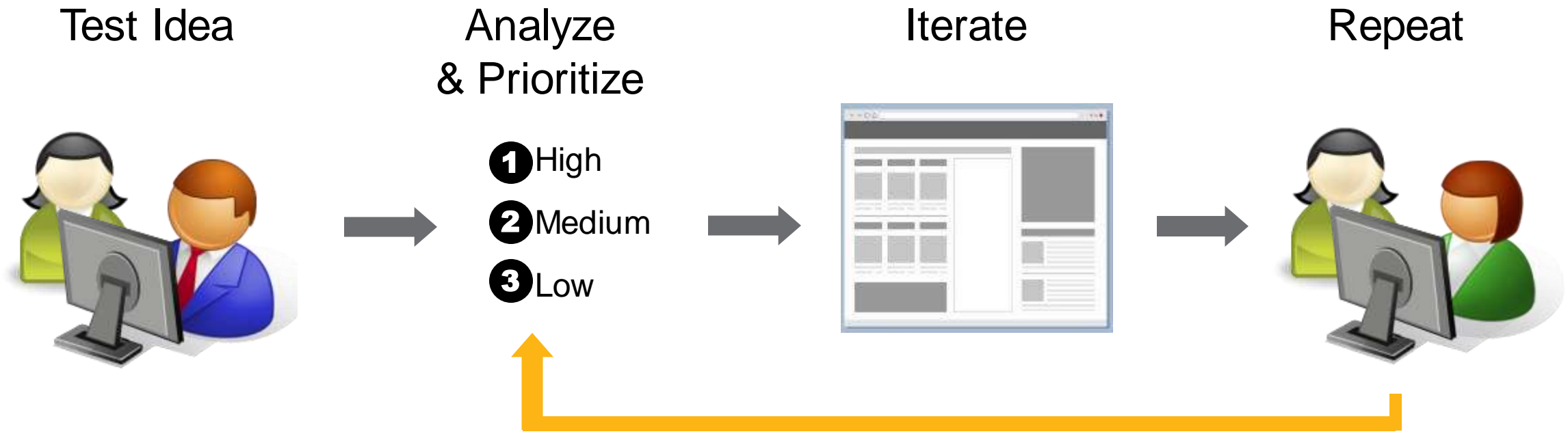


# Prototype – Experiments to Learn

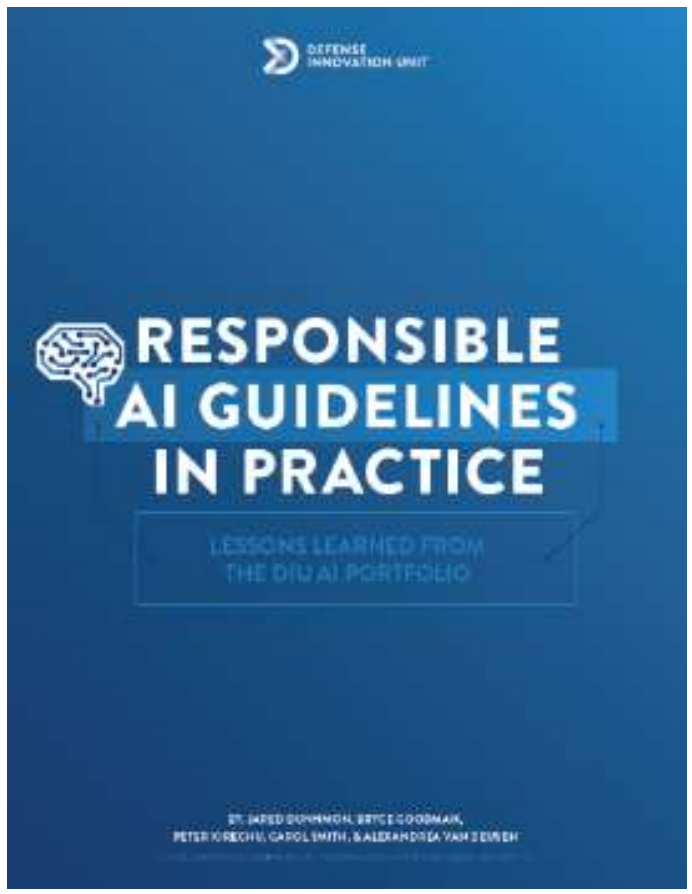
- How can I best support the user?
- How do I solve this problem?
- What interaction will meet the end users' needs?



# Iterative Cycles: Feedback and Improvement



# Presentations and Publications



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/1910.03515>  
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



What is a tomato?

Fruit?

Vegetable?

# Computer Vision - Image Recognition

## Train set



## Data encountered



Use case courtesy of Dr. Eric Heim, CMU Software Engineering Institute

# Only know what taught

## Train set



Unrepresentative or incomplete training data

## Data encountered



Unlikely to recognize

“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>

THE  
OPEN MIND

# Data is biased

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

	Gender	Age*	Detected
IBM	M	21**	✓
Microsoft	✗	✗	✗
Face++	✗	✗	✗
Kairos	M	29	✓

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

# Algorithms of Oppression, Dr. Safiya Noble

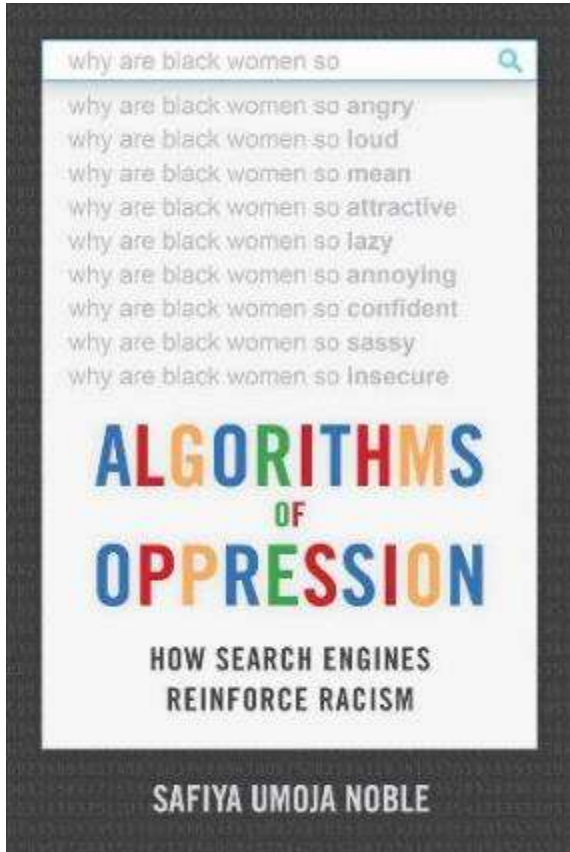


Photo from article: Google This: Algorithmic Oppression. ACLU News & Commentary. February 19, 2021. <https://www.aclu.org/news/privacy-technology/google-this-algorithmic-oppression/>



Responsible,  
Intentional  
Design

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

# 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, timely, and usable artificial intelligence (AI) systems with a diverse, team-aligned, trusted ethics. An initial version of this document was presented with the paper *Designing Trustworthy AI: A Human-Machine Learning Framework for Safe Development* by Carol Smith available at <https://arxiv.org/abs/1910.03515>.

<p><b>We will design our AI system with the following in mind:</b></p> <ul style="list-style-type: none"> <li>□ Designers humans bear the ultimate responsibility for all decisions and outcomes:           <ul style="list-style-type: none"> <li>• Responsibilites are explicitly defined between the AI system and humans, and how they are shared.</li> </ul> </li> <li>• Human responsibility will be preserved for final decisions that affect a persons life, quality of life, health, or reputation.</li> <li>• Humans are always able to monitor, control, and deactivate systems.</li> </ul> <p>□ Significant decisions made by the AI system will be:           <ul style="list-style-type: none"> <li>• explainable</li> <li>• auditable and verifiable</li> <li>• portable and reversible</li> </ul> </p>	<p><b>We will, or specifically identify the full range of risks and benefits:</b></p> <ul style="list-style-type: none"> <li>□ Harmful, malicious use and consequences, as well as good, beneficial use and consequences.</li> <li>□ We will be cognizant and exclusively research foreseeable consequences.</li> </ul> <p><b>We will create plans for the manufacture of the AI system, including the following:</b></p> <ul style="list-style-type: none"> <li>□ Communication plans to keep payment information with affected people</li> <li>□ Mitigation plans for managing the identified speculative risks</li> </ul> <p><b>We value respect and security:</b></p> <ul style="list-style-type: none"> <li>□ Incorporating our values of humanity, ethics, equity, fairness, accessibility, diversity and inclusion.</li> <li>□ respecting privacy and data rights (only necessary data will be collected)</li> <li>□ providing understandable security warnings</li> <li>□ making the AI system secure, valid, and reliable</li> </ul>	<p><b>We value transparency with the goal of engendering trust:</b></p> <ul style="list-style-type: none"> <li>□ The purpose, limitations, and biases of the AI system are explained in plain language.</li> <li>□ Designers track, log, and manage known and expected risks and known and expected stated algorithmic and modelic shortcomings and failures.</li> <li>□ Confidence and competence preserved for humans in decision making.</li> <li>□ Transparent justification for recommendations and outcomes is provided.</li> <li>□ Single password and interpretable monitoring systems are available.</li> </ul> <p><b>We value honesty and usability:</b></p> <ul style="list-style-type: none"> <li>□ Humans are readily able to enter they are interacting with the system - i.e. human</li> <li>□ Humans are readily directed when and why the AI system is taking action and/or making decisions.</li> <li>□ Engagements will be made regularly to meet human needs and technical standards.</li> </ul>
--	---	--

Team Signatures and Date:

---

**About the SEI**  
The Software Engineering Institute (SEI) is a non-profit organization that provides research and education in software engineering. SEI is a part of Carnegie Mellon University and is located on the campus of Carnegie Mellon University in Pittsburgh, Pennsylvania. SEI is a leader in the field of software engineering and has been recognized for its contributions to the field for many years.

**Contact Us**  
1-800-393-7339  
 3615 University City Drive, Pittsburgh, PA 15261-1501  
 412-263-1000  
 412-263-1001  
 412-263-1002

© 2023 Carnegie Mellon University. All rights reserved. CMU-SEI-2023-001

# 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, IBMer  
[https://www.ted.com/talks/grady\\_booch\\_don\\_t\\_fear\\_superintelligence](https://www.ted.com/talks/grady_booch_don_t_fear_superintelligence)

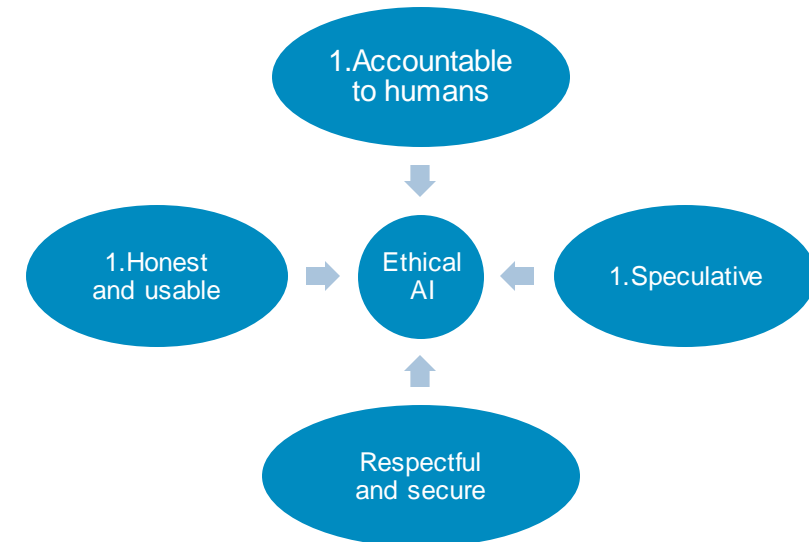
# Reward team members for finding ethics bugs

**Ayanna  
Howard**



# 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

# 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:

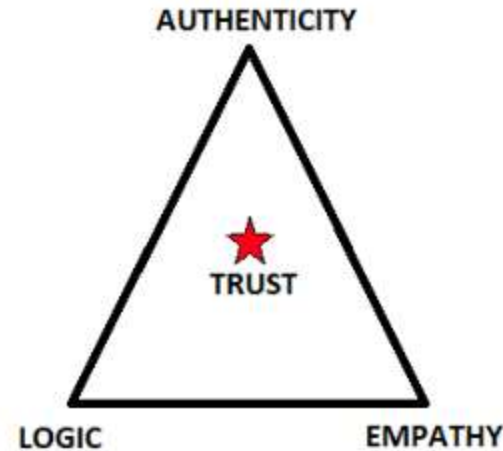
- push you out of your comfort zone
- have different levels of experience than you
- are from a different background
- are good at things you're terrible at

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

Carol J. Smith

LinkedIn: <https://www.linkedin.com/in/caroljsmith/>

