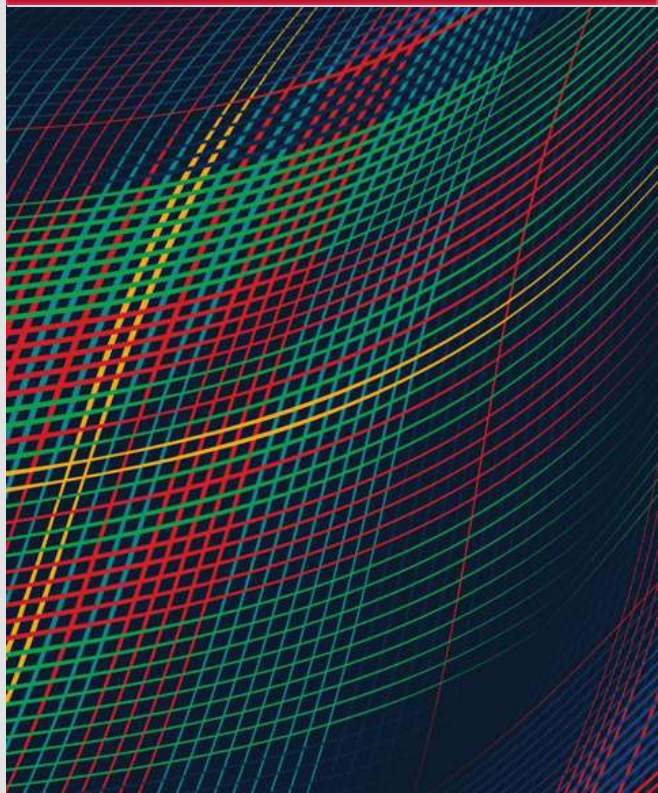
Carnegie Mellon University
Software Engineering Institute



Ethics and Trust for Emerging Technologies

NTIR 2023 | UALBANY CEHC

Carol J. Smith

Sr. Research Scientist, Human-Machine Interaction, AI Division
Adjunct Instructor, Human-Computer Interaction Institute

Copyright Statement

DM23-0428

About ACM

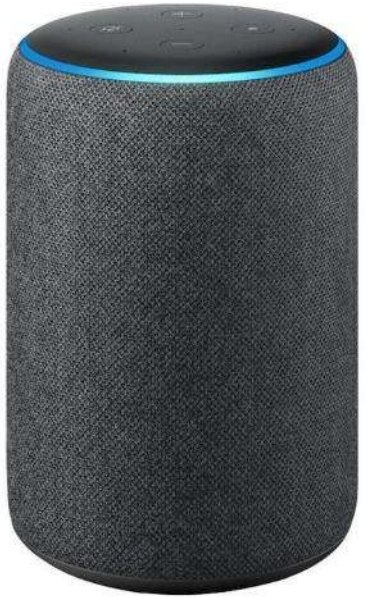
Carnegie Mellon University Software Engineering Institute

- ACM, the Association for Computing Machinery (www.acm.org), is the premier global community of computing professionals and students with nearly 100,000 members in more than 170 countries interacting with more than 2 million computing professionals worldwide.
- OUR MISSION: We help computing professionals to be theirs best and most creative. We connect them to their peers, to what the latest developments, and inspire them to advance the profession and make a positive impact on society.
- OUR VISION: We see a world where computing helps solve tomorrow's problems – where we use our knowledge and skills to advance the computing profession and make a positive social impact throughout the world.
- I am proud to be an ACM Member.

The Distinguished Speakers Program is made possible by:

For additional information, please visit http://speakers.acm.org

Emerging technology



Great potential - develop with caution

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.

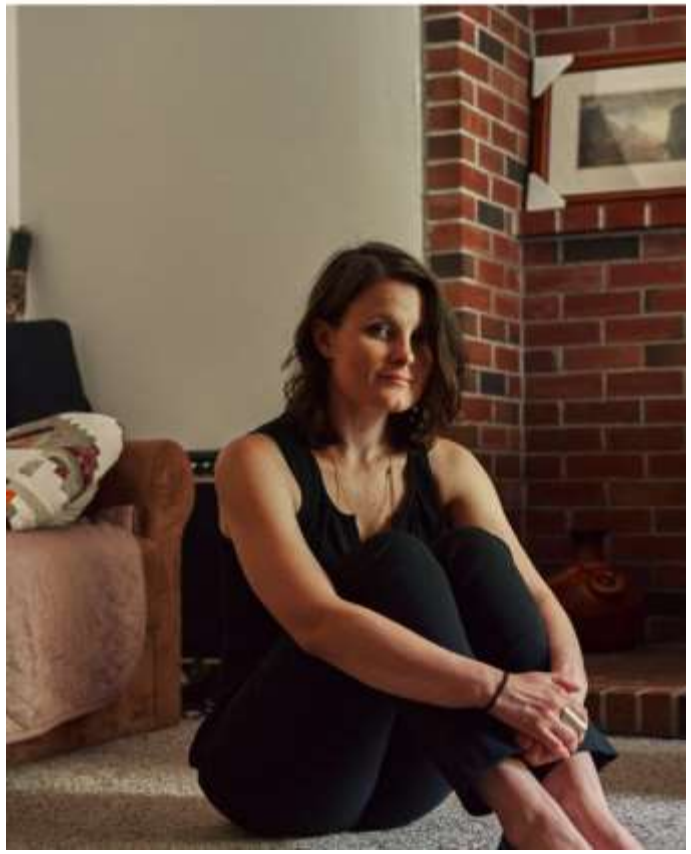
By **Mark Hanrahan**
December 12, 2019, 9:56 PM • 7 min read

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



Responsible, Intentional Design

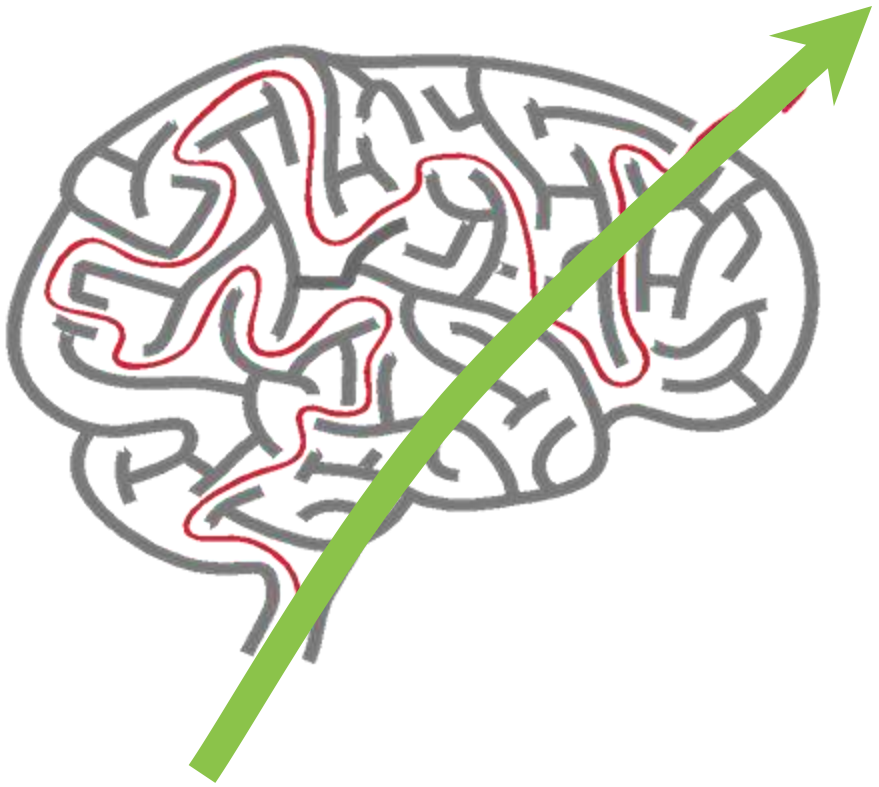
Carnegie Mellon University
Software Engineering Institute

Ethics

- Based on well-founded standards of right and wrong
- Standard of expected behavior that guides the correct course of action
- What impact does my work have?

What is Ethics? By Manuel Velasquez, Claire Andre, Thomas Shanks, S.J., and Michael J. Meyer. Markkula Center for Applied Ethics
https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/what-is-ethics/

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**

All systems have some form of bias

- Data is collected/curated, by humans, for a purpose.

- Complete objectivity is misleading.
- Bias can have purpose and can be helpful.

- Our Goal: **Reduce unintended and/or harmful bias**.

Image Recognition

Carnegie Mellon University Software Engineering Institute

Train set

Data encountered

Use case courtesy of Dr. Eric Heim, CMU SEI
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

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

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

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

Transparency and accountability

Carnegie Mellon University Software Engineering Institute

Understand inherent bias and amount of variance in the data:

- Creator's motivation
- Collection process
- Data included, and excluded
- Recommended uses, etc.

Provide evidence as appropriate

High value in diverse teams

Carnegie Mellon University Software Engineering Institute

Diverse teams

- focus more on facts
- process facts more carefully
- are more innovative

"…become more aware of their own potential biases"

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

David Rock, Heidi Grant. 2019. Why Diverse Teams Are Smarter. *Harvard Business Review*. November 4, 2019. https://hbr.org/2016/11/why-diverse-teams-are-smarter

Support collaborative work

- Identify biases within ourselves and our teams
- Introduce tools and processes to control for bias
- Be aware of biases emerging in our work

Early, purposeful work

In addition to understanding people and context

- Are we adding unnecessary additional risk to the situation?
- How will the system partner with people?
- How might these systems be misused/abused?

Does trust really matter?

Carnegie Mellon University
Software Engineering Institute

Trust

Trust is complex, transient, and personal.

Trust is a psychological state

- Use evidence to determine risk.
- Gained with enough confidence in positive outcomes, to give control of something significant, to the system.

Trust is contextual

Carnegie Mellon University Software Engineering Institute

Calibrated based on personal experiences, current context, and the available evidence of the system's capability and integrity.

Distrust
Trust falling short of system capabilities - may lead to disuse.

Calibrated Trust
Trust matches system capabilities - leading to appropriate use.

Over Trust
Trust exceeding system capabilities - may lead to misuse.

Rejection.

Automation bias.

John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. Hum Factors 46, 1 (March 2004), 50–80. DOI:https://doi.org/10.1518/hfes.46.1.50_30392
Bobbie Seppelt and John Lee. 2012. Human Factors and Ergonomics in Automation Design. In Handbook of Human Factors and Ergonomics (Fourth Edition) Chapter 59. Wiley. DOI: https://doi.org/10.1002/9781118131350.ch59

Trust depends on psychological safety

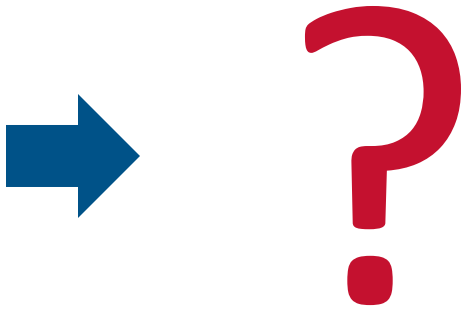
UX Framework

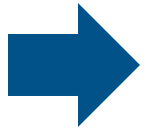
Implementing Ethics

Carnegie Mellon University Software Engineering Institute

How do we make tech human-centered?

User Experience Honeycomb
Peter Morville, et al.





Human-Centered Technology

Design to work with, and for, people

User Experience/Human-Computer Interaction research to understand:

- Complexity of current context (environmental, human, and via information)
- System requirements
- How system will need to respond to change
- Overall changes, over time and experience

Speculation keeps people safe - Activate Curiosity

Make Safe Experiences

Carnegie Mellon University Software Engineering Institute

Actions to get into or maintain a **safe state** should be **easy** to do.

Actions that can lead to an **unsafe state** (hazard) should be **hard** to do.

Be speculative
- Don't assume only average cases
- Can't verify probabilities

N. G. Leveson. 2017. The Therac-25: 30 Years Later. In Computer, vol. 50, no. 11, (November 2017), 8-11. DOI: 10.1109/MC.2017.4041349
N. Leveson. 1995. Safeware: System Safety and Computers, Addison Wesley (1995).

Capitalize on Human Strengths

Carnegie Mellon University Software Engineering Institute

- Humans are (still) better at many activities:

- Exposing Bias
- Identifying downstream impacts
- Judgment
- Recognizing Bias
- Responding to change
- Socio-political nuance
- Taking context into consideration

Amanda Muller and Carol Smith. 2022. Perceptions of Function Allocation between Humans and AI-Enabled Systems. UXPA 2022 (pre-print). https://uxpa2022.org/sessions/perceptions-of-function-allocation-between-humans-and-ai-enabled-systems/

Adopt Technology Ethics

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

Association for Computing Machinery

AINOW INSTITUTE

Microsoft

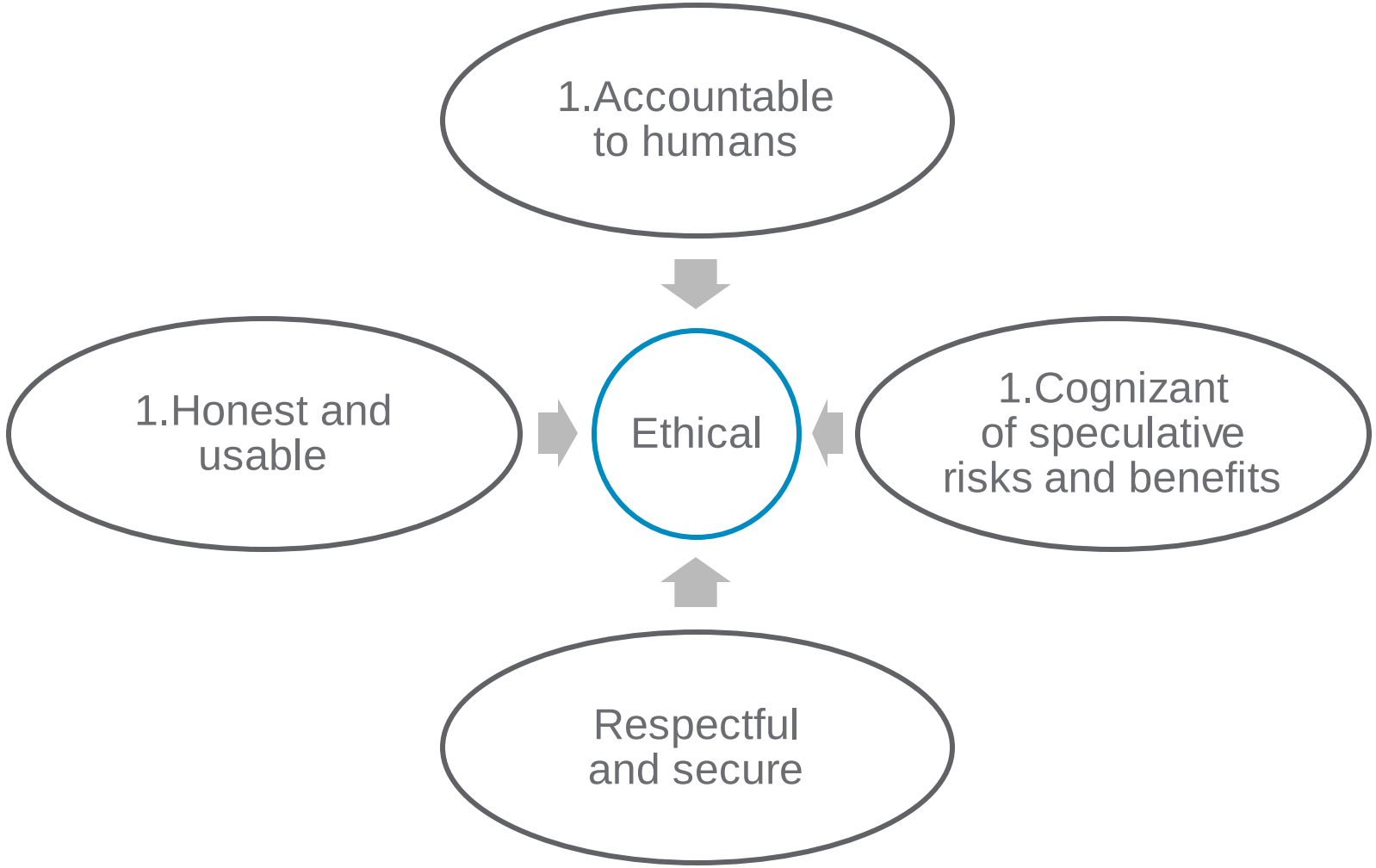
Google

Montréal Declaration Responsible AI_

An initiative of Université de Montréal

U.S. DEPT OF DEFENSE

Prompt conversations – UX Framework

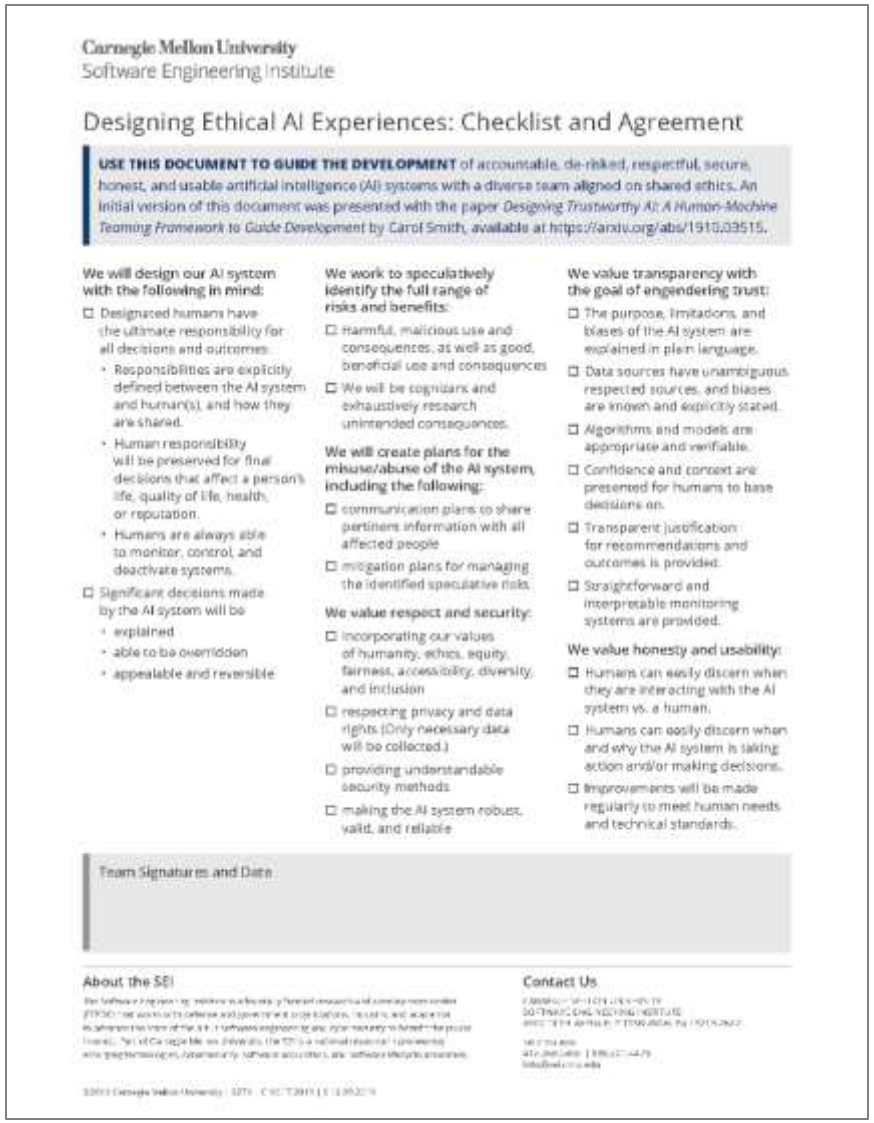


Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020. https://insights.sei.cmu.edu/sei_blog/2020/03/designing-trustworthy-ai-for-human-machine-teaming.html

Prompts for conversations

Pair checklists with technical ethics

- Bridge gaps 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

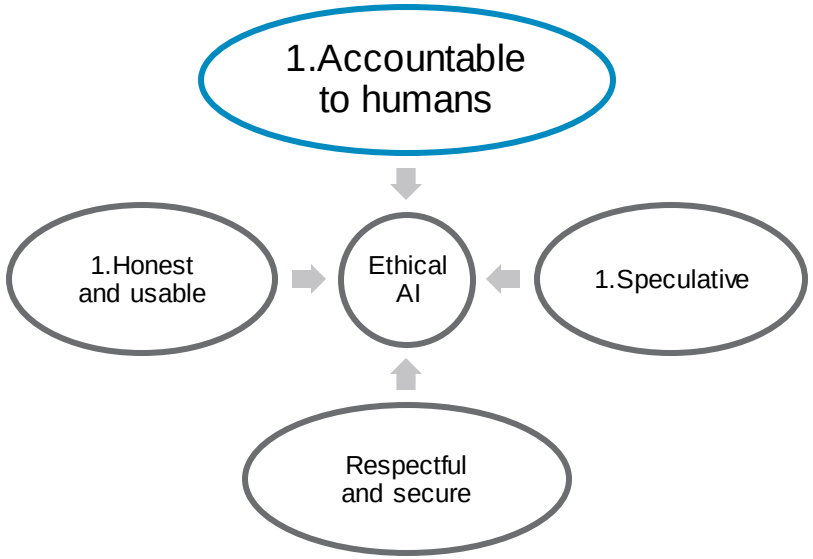
Contact Us

Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020.
Checklist and Agreement - Downloadable PDF: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620

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



Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020.

“Ensure humans can unplug the machines”

– Grady Booch

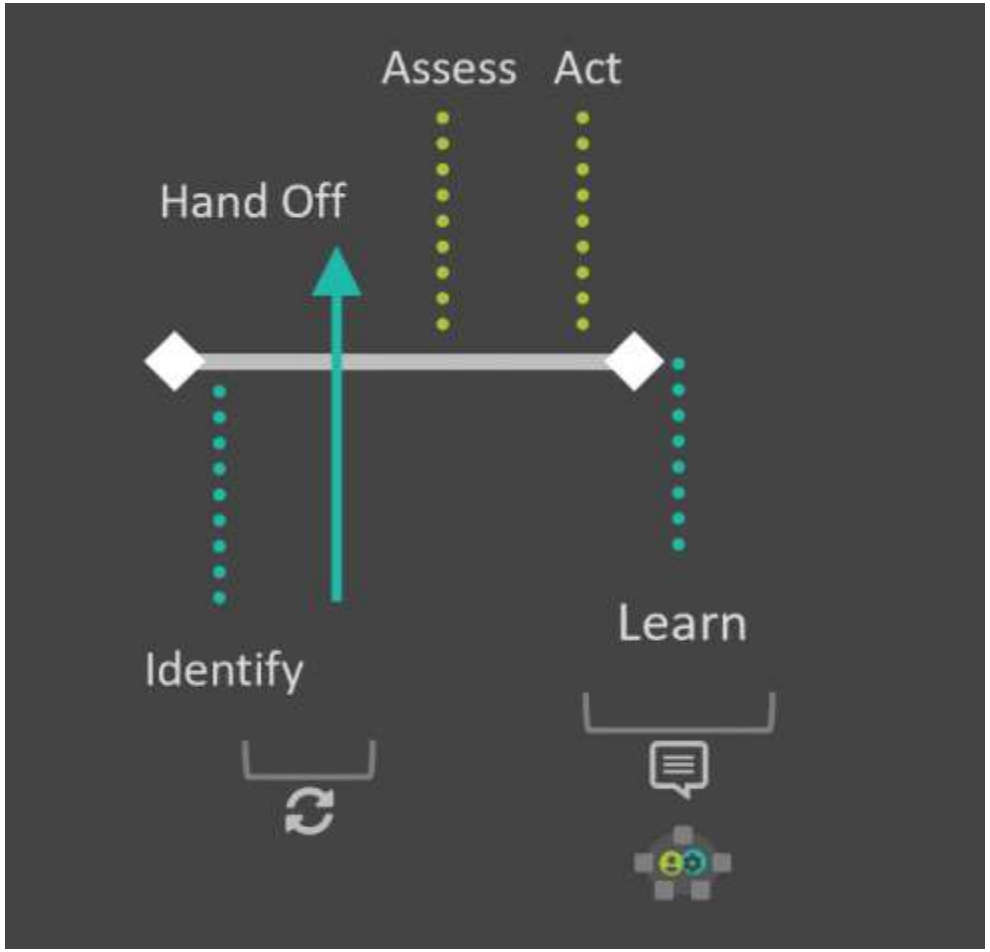
TED Talk, Grady Booch, Scientist, Philosopher, IBM’er

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

Significant decisions

Significant decisions made by system
- explained
- able to be overridden
- appealable and reversible

Responsibilities explicitly defined between people and systems.



Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020.

How IAs Can Shape the Future of Human-AI Collaboration. Carol Smith and Duane Degler. Presented on April 28-30, 2021 at IAC21.

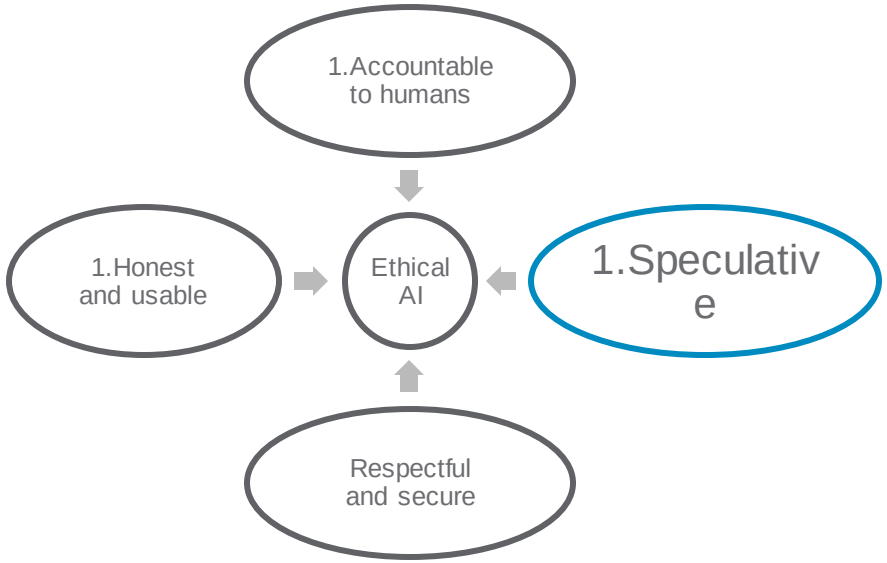
Cognizant of Speculative Risks and Benefits

Carnegie Mellon University Software Engineering Institute

Identify full range of

- Harmful, malicious use, as well as good, beneficial use.
- Unwanted/unintended consequences.

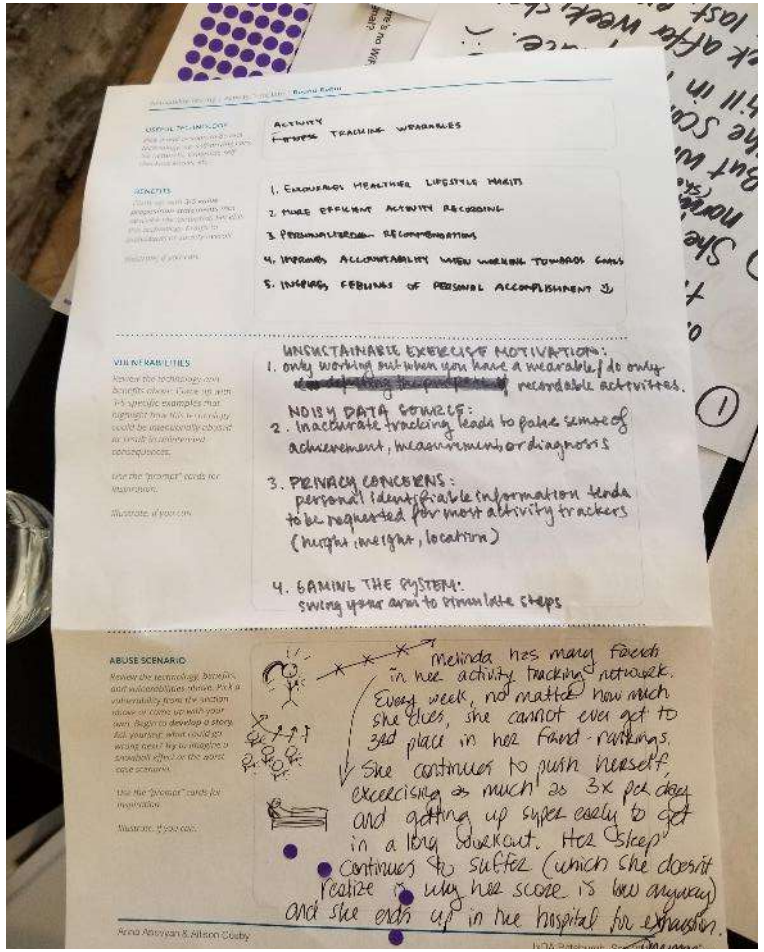
Prevent potential harms.



Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020.

Conduct UX research - activate curiosity

- Speculate about misuse and abuse – abusability testing
- Potentially severe consequences (even if rare)
- Perspective of people in frequently marginalized groups



Template by: Anna Abovyan & Allison Cosby, IxDA Pittsburgh, Sep 2019

Create communication & mitigation plans

Carnegie Mellon University Software Engineering Institute

Plan for unwanted consequences.

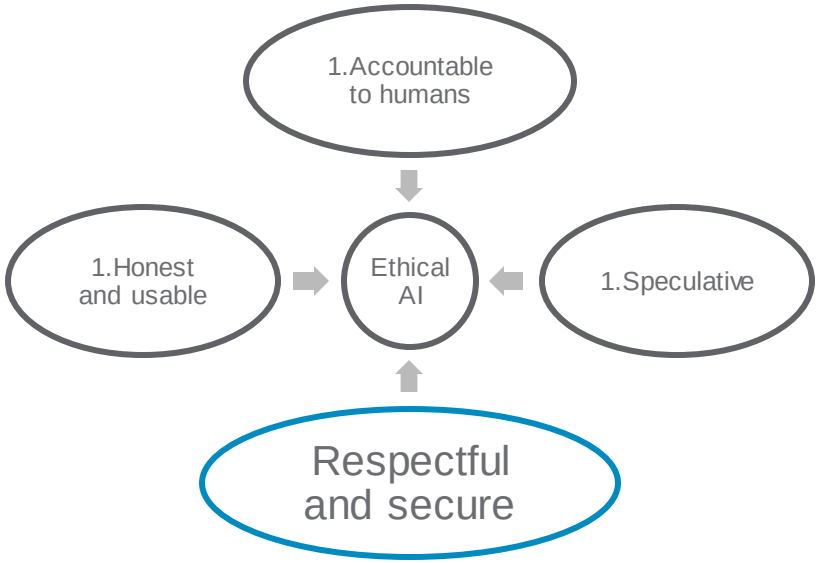
Misuse and abuse of system

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

Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020.

Respectful and Secure

- Values of humanity, ethics, equity, fairness, accessibility, diversity and inclusion.
- Respect privacy and data rights (only collect what is necessary).
- Make systems robust, valid, and reliable.
- Provide understandable security.



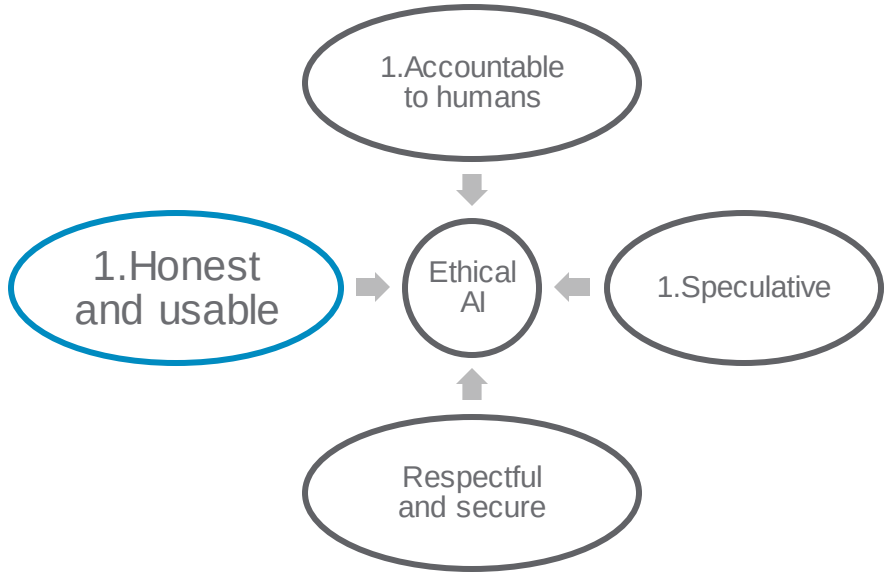
Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020.

Honest and Usable

- Value transparency with the goal of engendering calibrated trust.
- Provide transparency regarding boundaries and unfamiliar scenarios.
- Explicitly state identity as an AI system.

Fairness

- Show awareness of purposeful bias.
- Provide AI system limitations.
- Overcommunicate on issues.



Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020.

Honest?

Jason Allen's A.I.-generated work, "Théâtre D'opéra Spatial," took first place in the digital category at the Colorado State Fair.Credit...via Jason Allen. https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

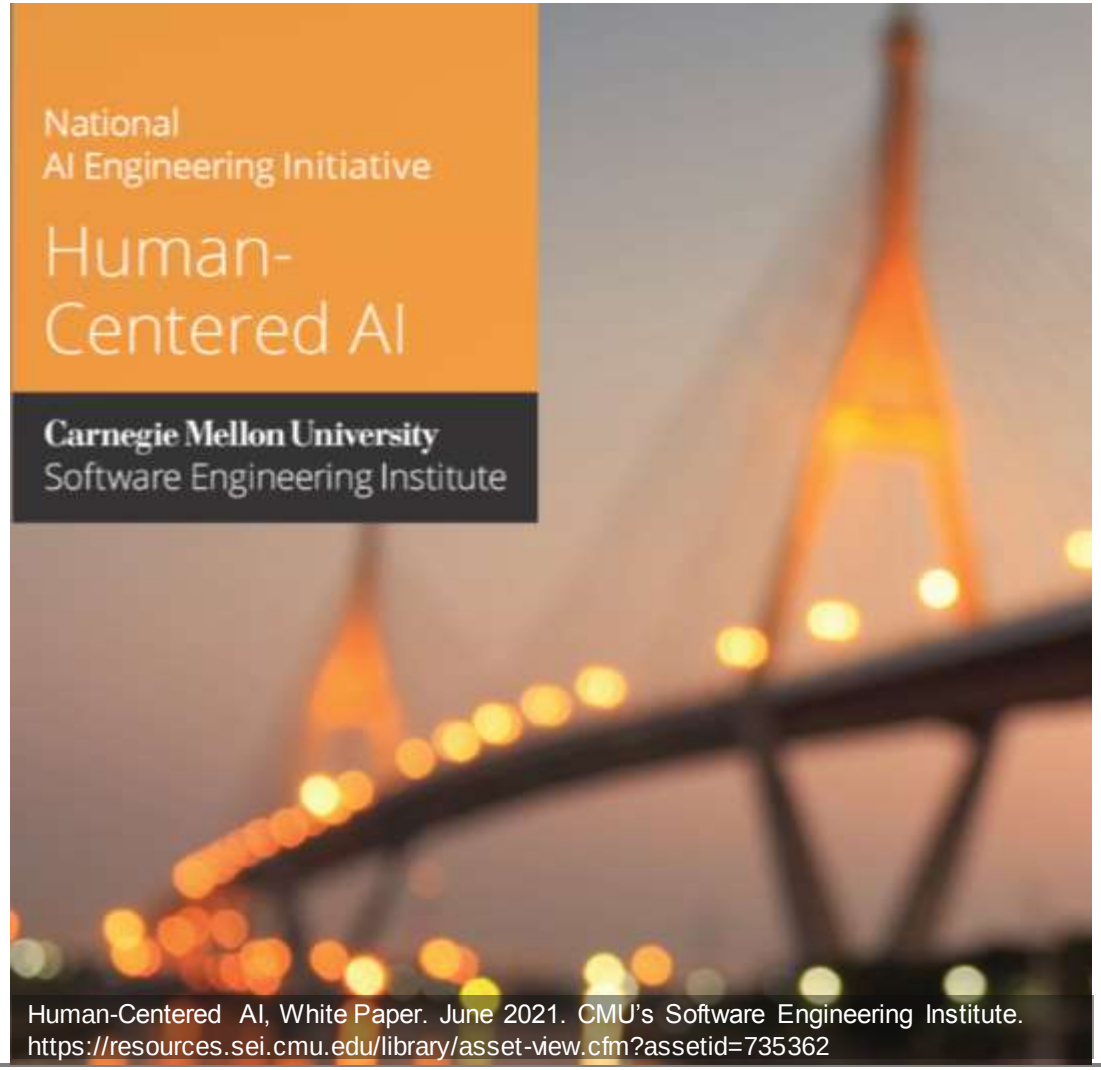
Design for Calibrated Trust

- Provide transparency regarding system limitations - boundaries and unfamiliar scenarios
- Speculate about misuse and abuse
- Prevent or plan to mitigate situation

Engage in critical oversight

"What are we doing?
Why are we doing it,
and for whom?"

- Continuous human oversight
- Identify risks of bias, misuse, abuse, and unintended consequences
- Proactively consider risks



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

Establish psychological safety for diverse teams

Change is constant

Conversations for Understanding

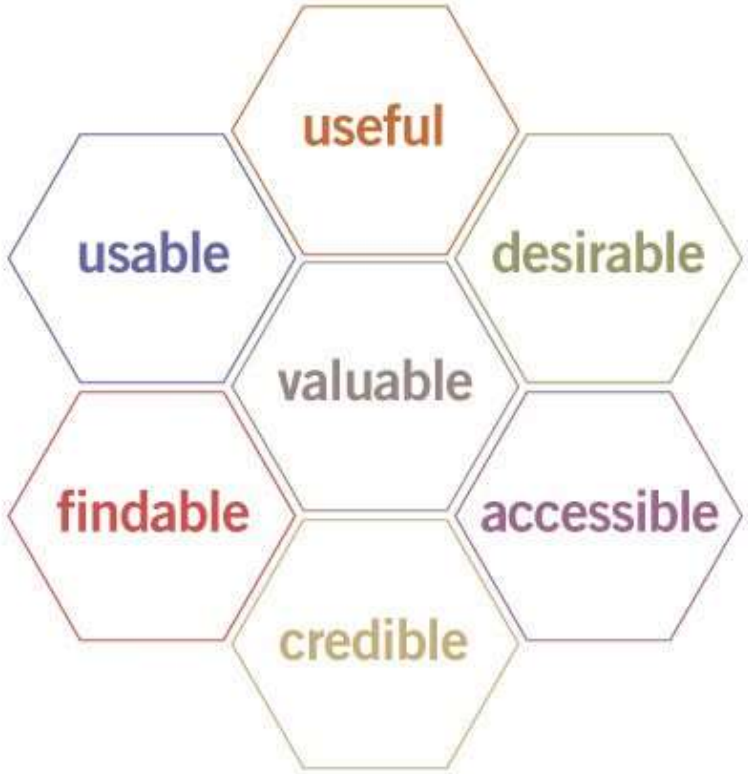
Difficult Topics

- What do we value?
- Who could be hurt?
- What lines won't our AI cross?
- How are we shifting power?*
- Perspective of frequently marginalized groups

*"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

Design to work with, and for, people



User Experience Honeycomb
Peter Morville, et al.

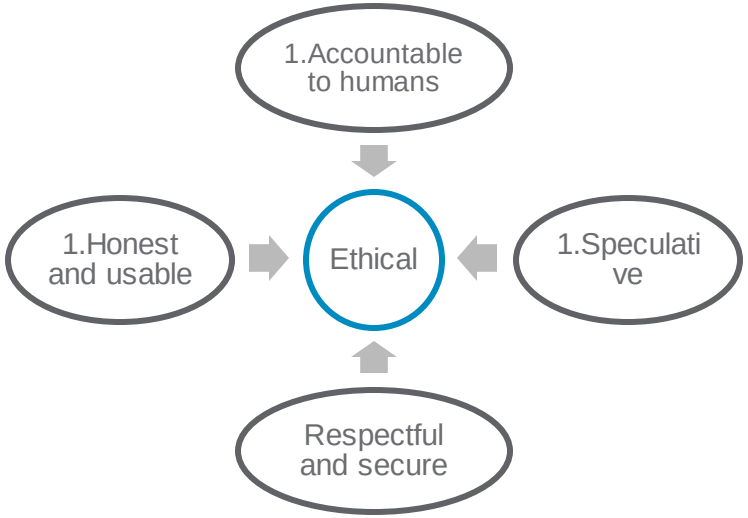
1. Capitalize on human strengths
2. Encourage deep conversations
3. Be intentional, keep people safe

Human-Centered Technology

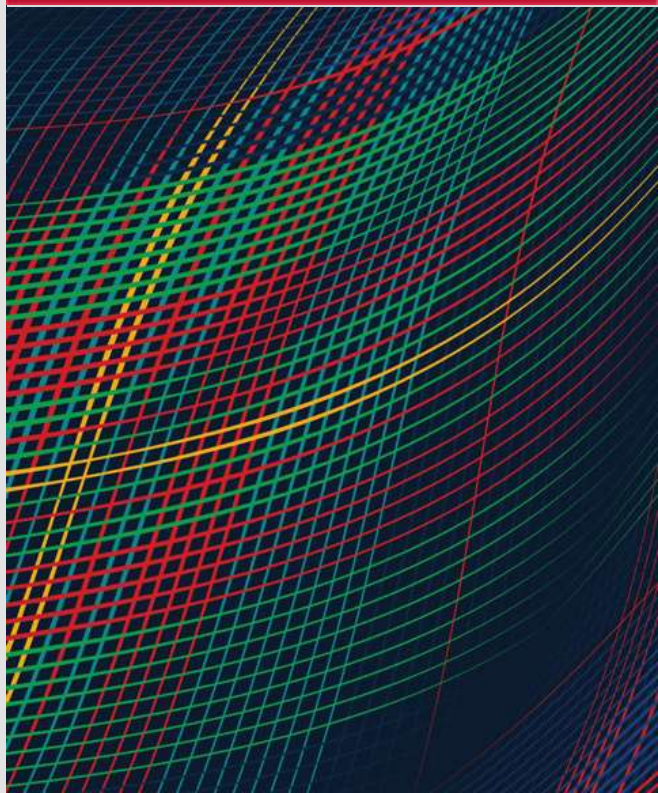
Carnegie Mellon University Software Engineering Institute

Be uncomfortable and kind.

We aren't perfect, tech won't be perfect.



Carnegie Mellon University
Software Engineering Institute



Carol J. Smith

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

AI DIVISION
SOFTWARE ENGINEERING INSTITUTE