Human-Centered and Responsible AI

CMU LISBON, PORTUGAL 2023

Carol J. Smith

Sr. Research Scientist, Human-Machine Interaction, AI Division

Adjunct Instructor, Human-Computer Interaction Institute

Carnegie Mellon University
Software Engineering Institute

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.

DM23-0538

Carol J. Smith

Carnegie Mellon University Software engineering institute

Software Engineering Institute

Human-Computer Interaction 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

First Machines

Al-Jazari described a water-powered automaton orchestra on a boat in 1206

But the lack of research results in…

Broaden our Work

Is this an AI-friendly challenge?
What kind of improvements are expected?
What are the benefits and risks?
How will we know we've made improvements?

AI is data

Carnegie Mellon University
Software Engineering Institute

Data

~90 hours of reading a week for physician to keep up with all published medical articles (2004)

AI could enable a physician to make more evidence-based decisions.

If data is available and AI *appropriate,*

Alper, Brian S. et al. “How Much Effort Is Needed to Keep up with the Literature Relevant for Primary Care?” Journal of the Medical Library Association 92.4 (2004): 429–437. Print.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC521514/

Computer Vision - Image Recognition

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

Train set

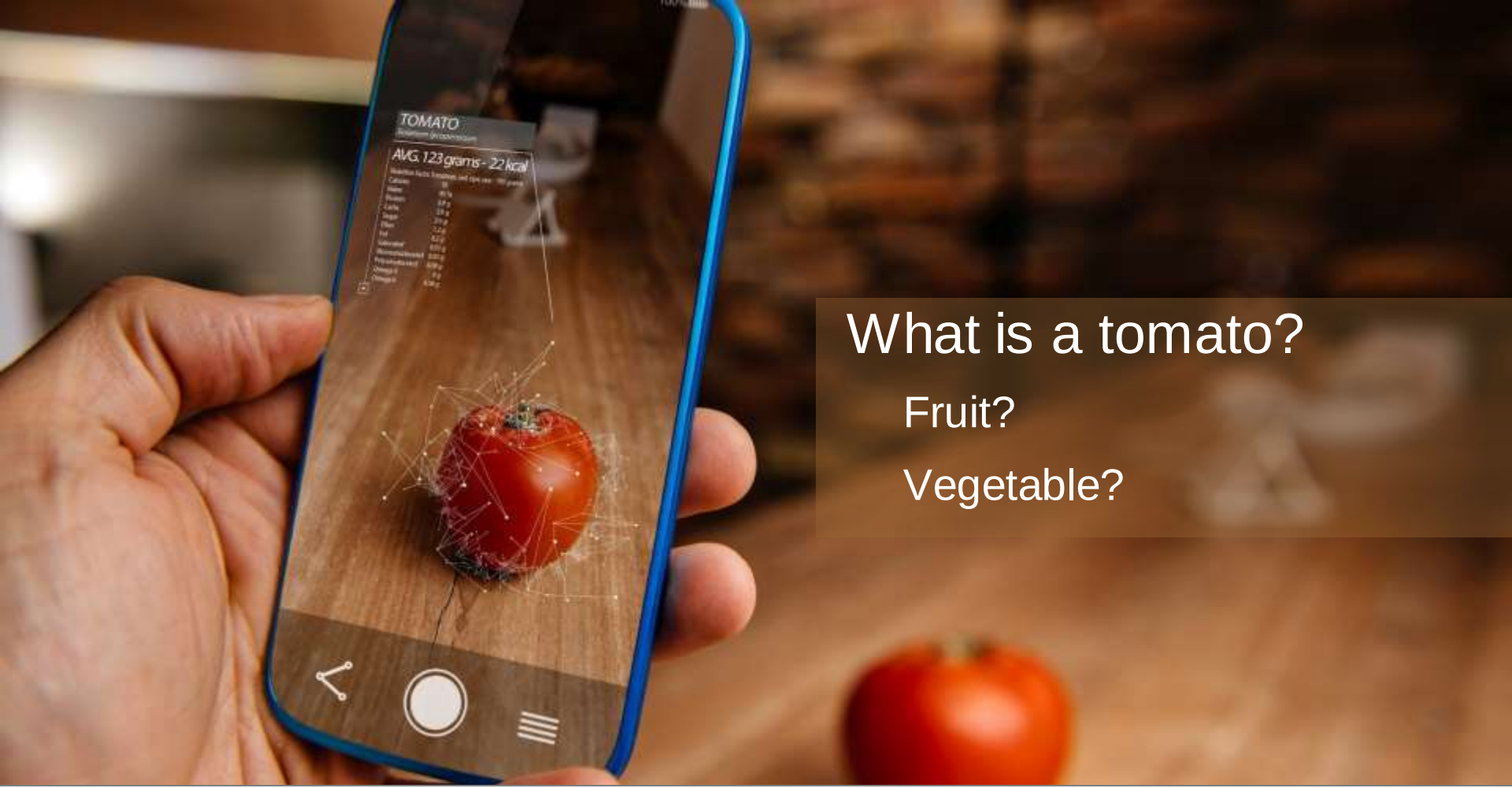
Unrepresentative
or incomplete training data

Data encountered

Unlikely to recognize

All systems have some form of bias

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



What is a tomato?

- Fruit?
- Vegetable?

Bias in data, algorithm selection, and training

Understand inherent bias and amount of variance

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

Goal: Reduce harmful bias

Avoid reinforcing discrimination against historically marginalized groups.

Removing all bias is impossible
- indicators are concealed in the data.

Removing obvious indicators (gender, zip code, etc.) reduces the ability to track bias.

"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

Sensing changes over time

Understand Complexity of Context

Carnegie Mellon University
Software Engineering Institute

Sources of Complexity

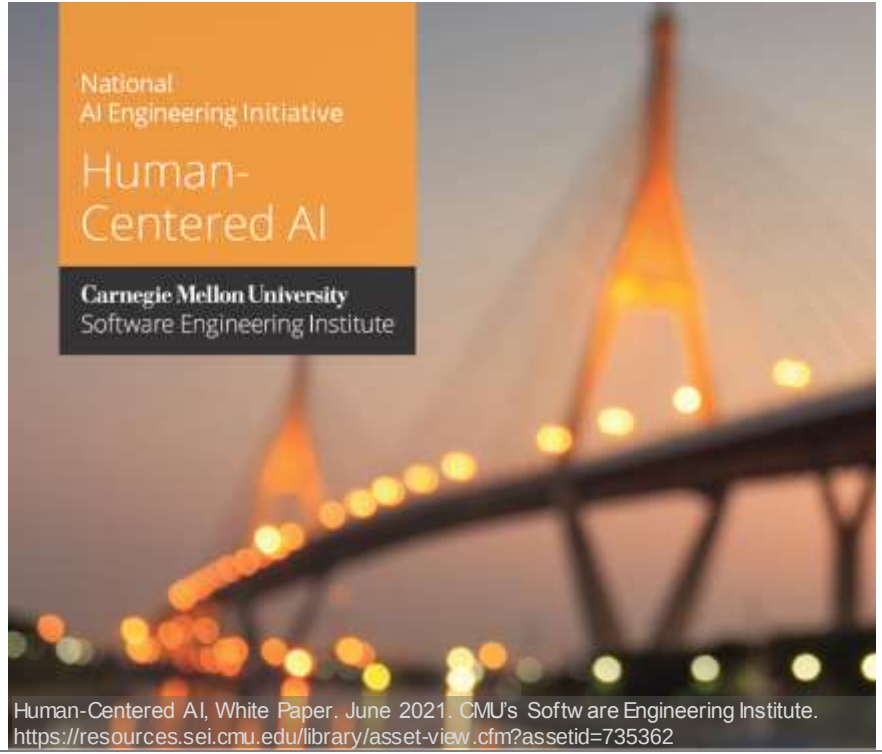
Environment
People
Information
System capabilities

Image by Alan Warburton / © BBC / Better Images of AI / Plant / CC-BY 4.0

Dynamic Contexts

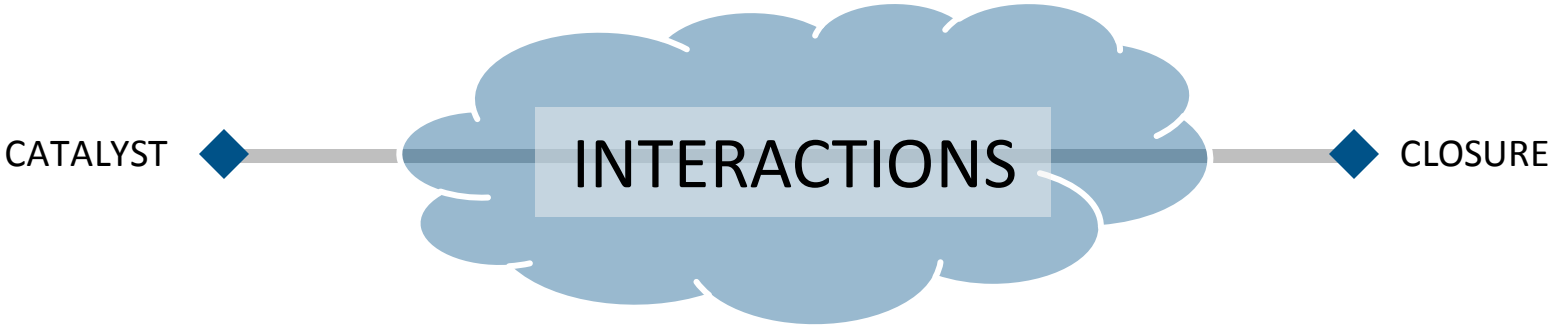
How do people and AI:

- learn when shifts in context have occurred?
- maintain clarity around operational intent?
- adapt and evolve based on dynamic contexts?



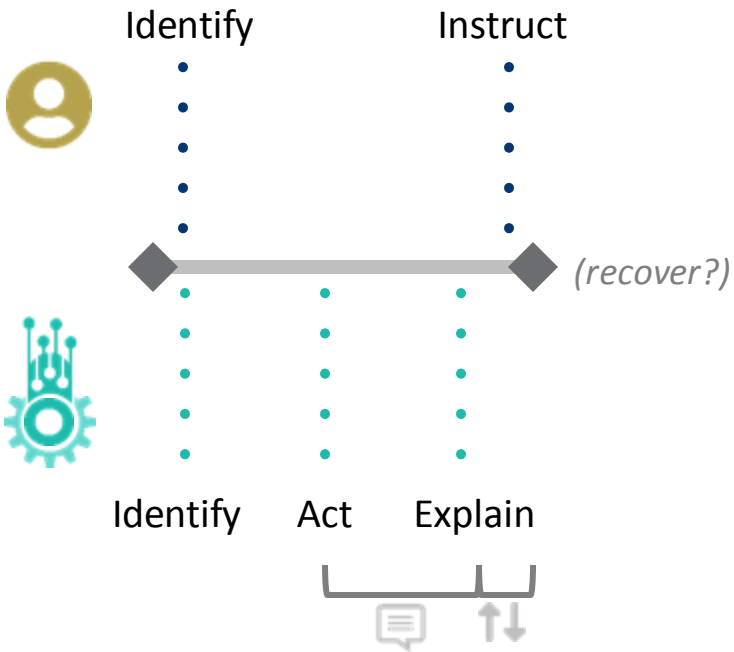
Human-Centered AI, White Paper. June 2021. CMU's Software Engineering Institute. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362

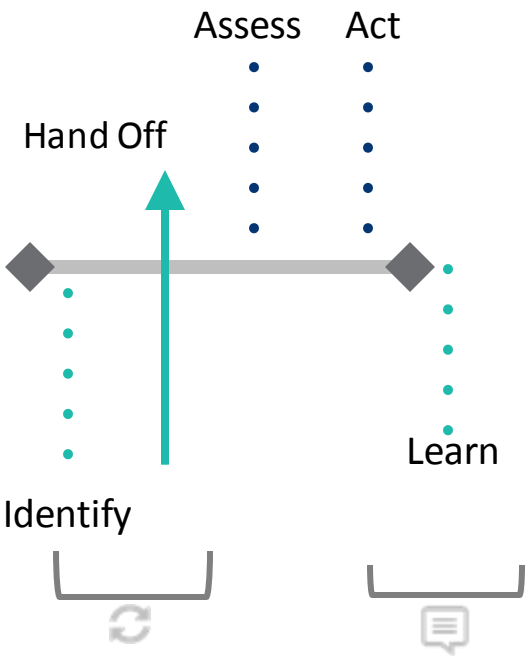
Exchange of information



How IAs Can Shape the Future of Human-AI Collaboration
Presented on April 28-30, 2021 at the Information Architecture Conference (IAC21)

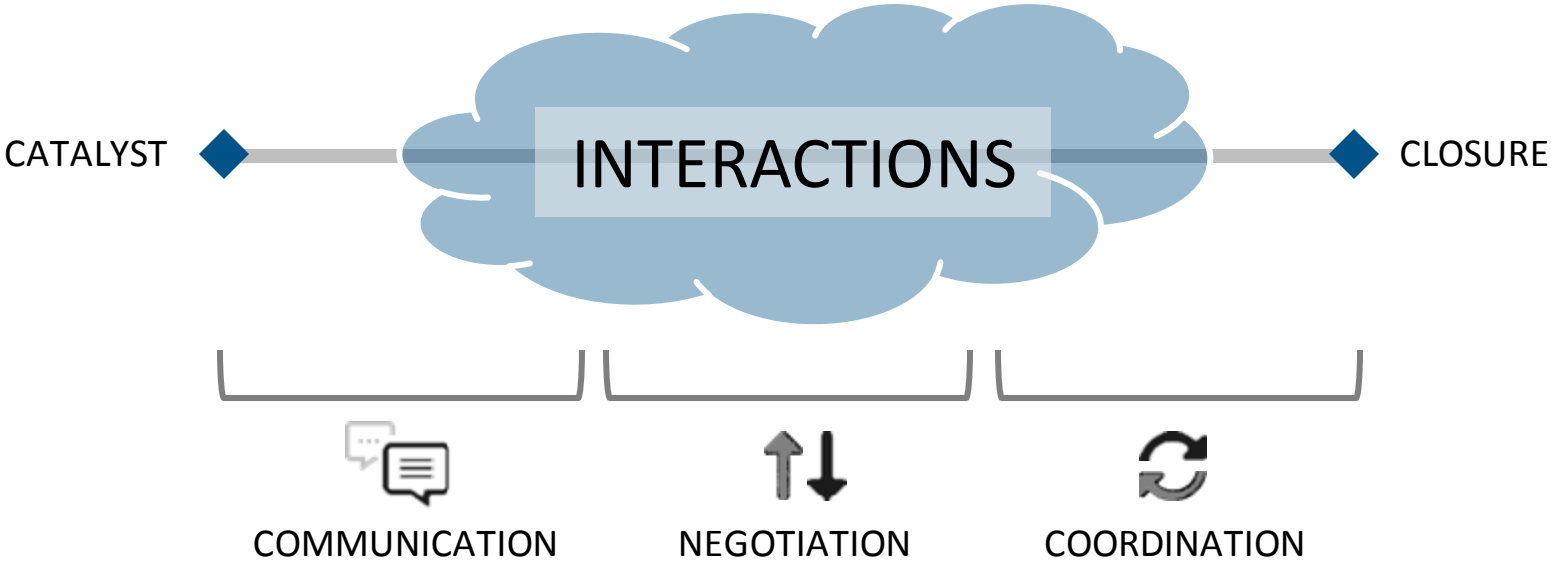
Semi-autonomous Vehicle Avoids Obstacle





How IAs Can Shape the Future of Human-AI Collaboration
Presented on April 28-30, 2021 at the Information Architecture Conference (IAC21)

Collaborative activities - interactions



How IAs Can Shape the Future of Human-AI Collaboration
Presented on April 28-30, 2021 at the Information Architecture Conference (IAC21)

Capitalize on Human Strengths

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/

Trustworthy AI

Carnegie Mellon University
Software Engineering Institute

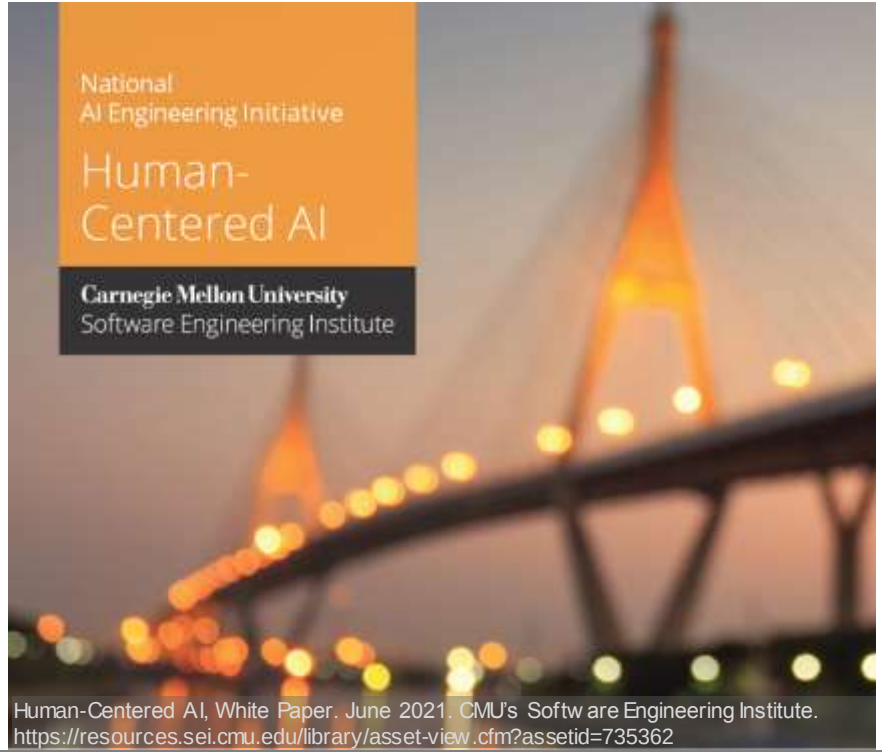
AI must be designed to work with, and for, people. People need to trust their tools to use them properly.

Trustworthy AI is Human-Centered

Capabilities (and limitations) are explained – transparency.

Continuous monitoring and oversight are prioritized.

People are enabled to gain *calibrated levels of trust*.



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

Provide Evidence

What is Calibrated Trust?

Trust is personal - a dynamic psychological state.
We calibrate trust based on personal experiences, current context, and available evidence of 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.

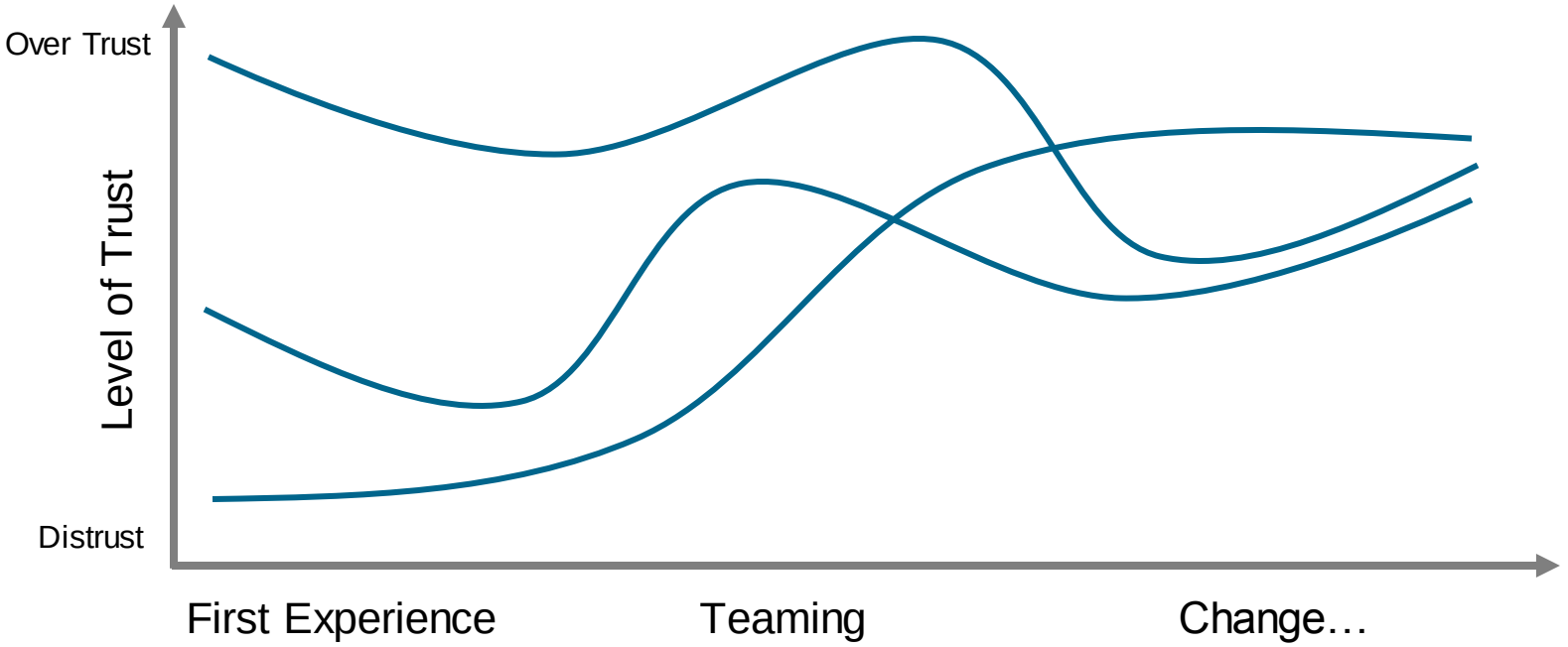
Over Trust
Trust exceeding system capabilities - may lead to misuse.

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: 10.1002/9781118131350.ch59
Mary Cummings. 2004. Automation Bias in Intelligent Time Critical Decision Support Systems. AIAA 2004-6313. AIAA 1st Intelligent Systems Technical Conference. (September 2004). DOI: 10.2514/6.2004-6313

Trust is Complex and Transient



Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User Trust Dynamics: An Investigation Driven by Differences in System Performance. IUI 2017 (March 2017), 307-317. DOI: http://dx.doi.org/10.1145/3025171.3025219

Broadening UX

Carnegie Mellon University
Software Engineering Institute

Broadening UX

User Experience Honeycomb
Peter Morville, et al.

Responsible
and
Human-Centered AI

Speculation Keeps People Safe

Activate Curiosity

Speculate about misuse and abuse

- Unintended and unwanted consequences
- Negative consequences for people who are frequently marginalized

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

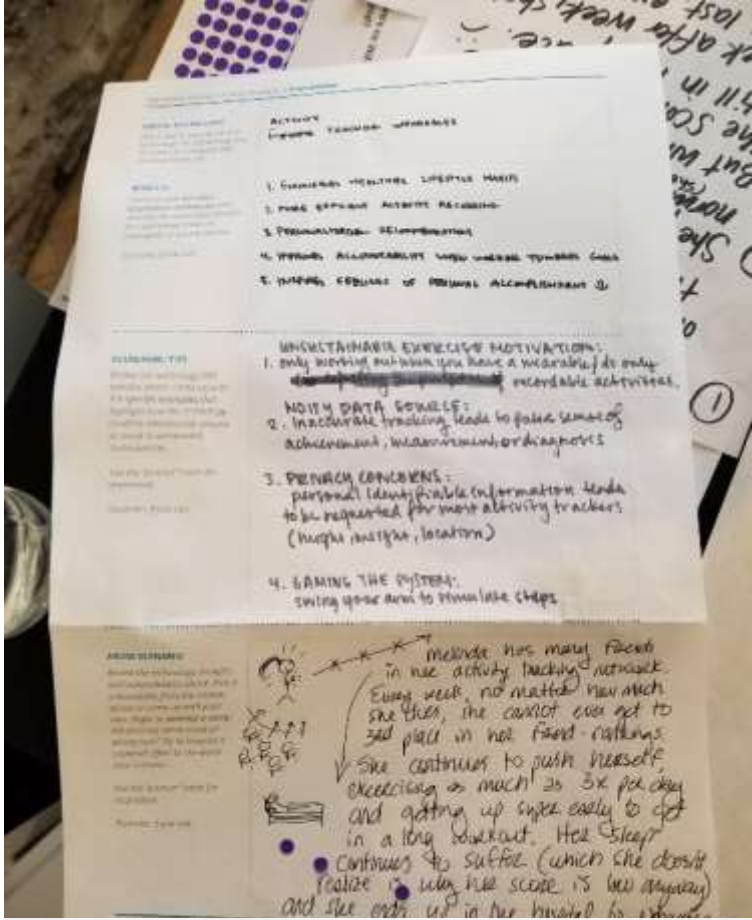
Abusability Testing

1) **Value proposition**
   Benefits tech brings to individuals, society

2) **Vulnerabilities**
   How tech could be misused

3) **Abuse scenario**

Provocation via prompt statements



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

UX in the Age of Abusability. The role of Composition, Collaboration, and Craft in building ethical products. Dan Brown. Sep 18, 2018. https://greenonions.com/ux-in-the-age-of-abusability-797cd01f6b13
Photo from workshop organized by Anna Abovyan, Theora Kvitka and Allison Cosby of the Pittsburgh IxDA Chapter for World Interaction Design Day 2019.

Reward team members for finding ethics bugs

Card Game: What Could Go Wrong?

Foster conversations around potential challenges and issues with complex technologies.

Figure 1: What Could Go Wrong? A digital card game to foster conversation around potential challenges and issues of autonomous vehicle technology.

Nikolas Martelaro and Wendy Ju. 2020. What Could Go Wrong? Exploring the Downsides of Autonomous Vehicles. In 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '20). Association for Computing Machinery, New York, NY, USA, 99–101. https://doi.org/10.1145/3409251.3411734

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

Photo by Pam Sharpe On Unsplash
https://unsplash.com/@msgrace?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText

Leaders establish psychological safety

Adopt Technology Ethics

Harmonize cultural variations.

Balance to pace of change.

Explicit permission to consider and question breadth of implications.

Association for Computing Machinery

AI NOW INSTITUTE

Microsoft

Google

Montréal Declaration Responsible AI

An initiative of Université de Montréal

U.S. DEPT OF DEFENSE

Prompt conversations

Checklists, frameworks, and guidelines – pair with technical ethics.

- Bridge gaps between “do no harm” and reality
- 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, trustworthy, 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 humans, 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 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

Team Signatures and Date

About the SEI

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

Defense Innovation Unit. Artificial Intelligence Portfolio, Responsible AI Guidelines. https://www.diu.mil/responsible-ai-guidelines

New uncomfortable work

"***Be uncomfortable***"

- Laura Kalbag

Ethical design is not superficial.

Big Topics

Carnegie Mellon University
Software Engineering Institute

Authorship and Accountability

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

Large Language Models – "Spicy Autocomplete"

Carnegie Mellon University Software Engineering Institute

Generative AI – LLMs

- Text - ChatGPT, Bing Chatbot, Bard, etc.
- Image – DALL-E, Midjourney (using text labels)

Results

- Grammatically correct – believable text
- Fabricate data (even URLs) - "Hallucinations"

Controversy

- Privacy and intellectual property
- Automation bias

Significant Decisions

Made by system

- explained
- able to be overridden
- appealable and reversible

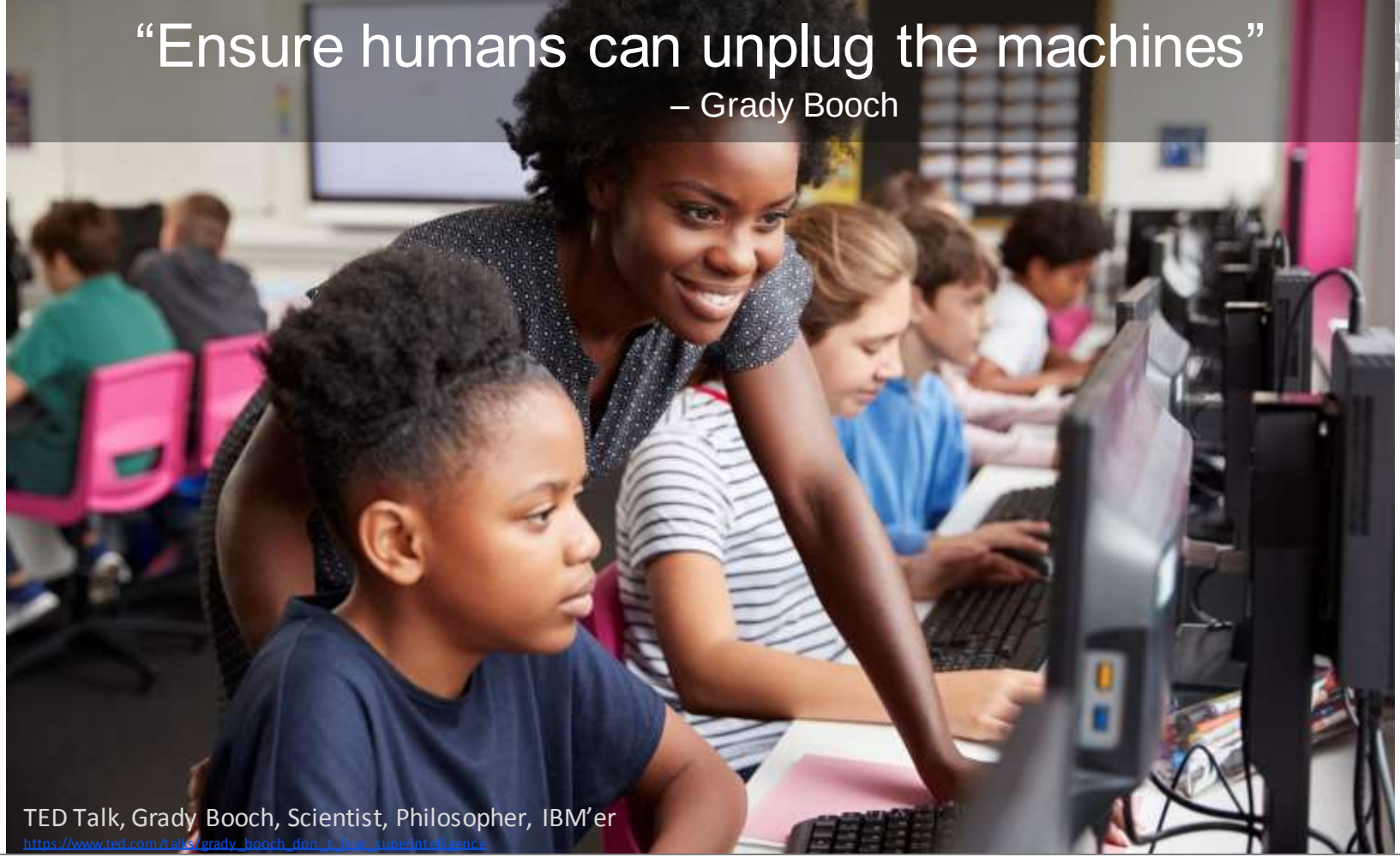
Responsibilities are explicitly defined between people and systems.

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.

“Ensure humans can unplug the machines”

– Grady Booch



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

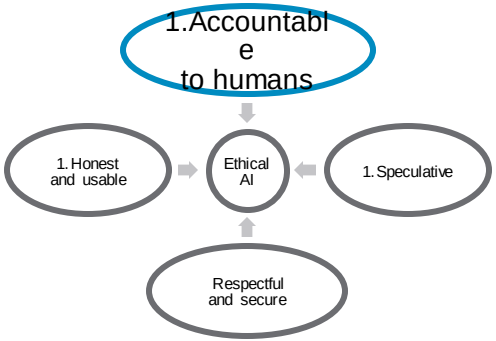
https://www.ted.com/talks/grady_booch_don_t_fear_superintelligence

Humans are Accountable

Ensure humans have ultimate control.
Able to monitor and control risk.

A person is always responsible 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.

Design AI to work with, and for, people

Carol J. Smith

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

AI DIVISION
SOFTWARE ENGINEERING INSTITUTE

Carnegie Mellon University
Software Engineering Institute