CMU HCII

UXR for Responsible, Human-Centered AI

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

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

[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.

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.

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

DM23-0085

Carol J. Smith

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

© 2023 Carnegie Mellon University

[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.

Efficiency and Information

© 2023 Carnegie Mellon University

[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.

Water-powered automaton orchestra on a boat, described by Al-Jazari in 1206.

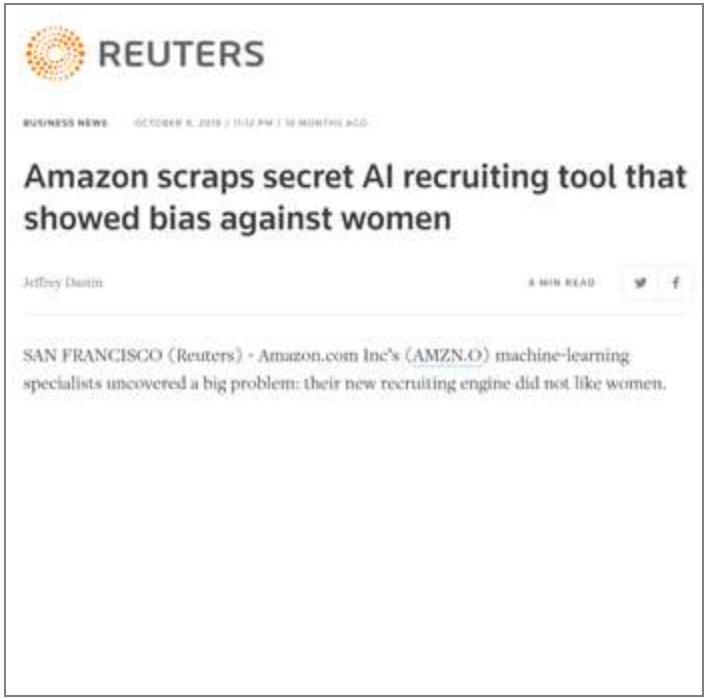
© 2023 Carnegie Mellon University

[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.

© 2023 Carnegie Mellon University

[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.

But the lack of research results in…



© 2023 Carnegie Mellon University

[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.

Make responsible and human-centered AI

User Experience Honeycomb
Peter Morville, et al.

Responsible
and
Human-Centered
(Ethical) AI

© 2023 Carnegie Mellon University

[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.

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?

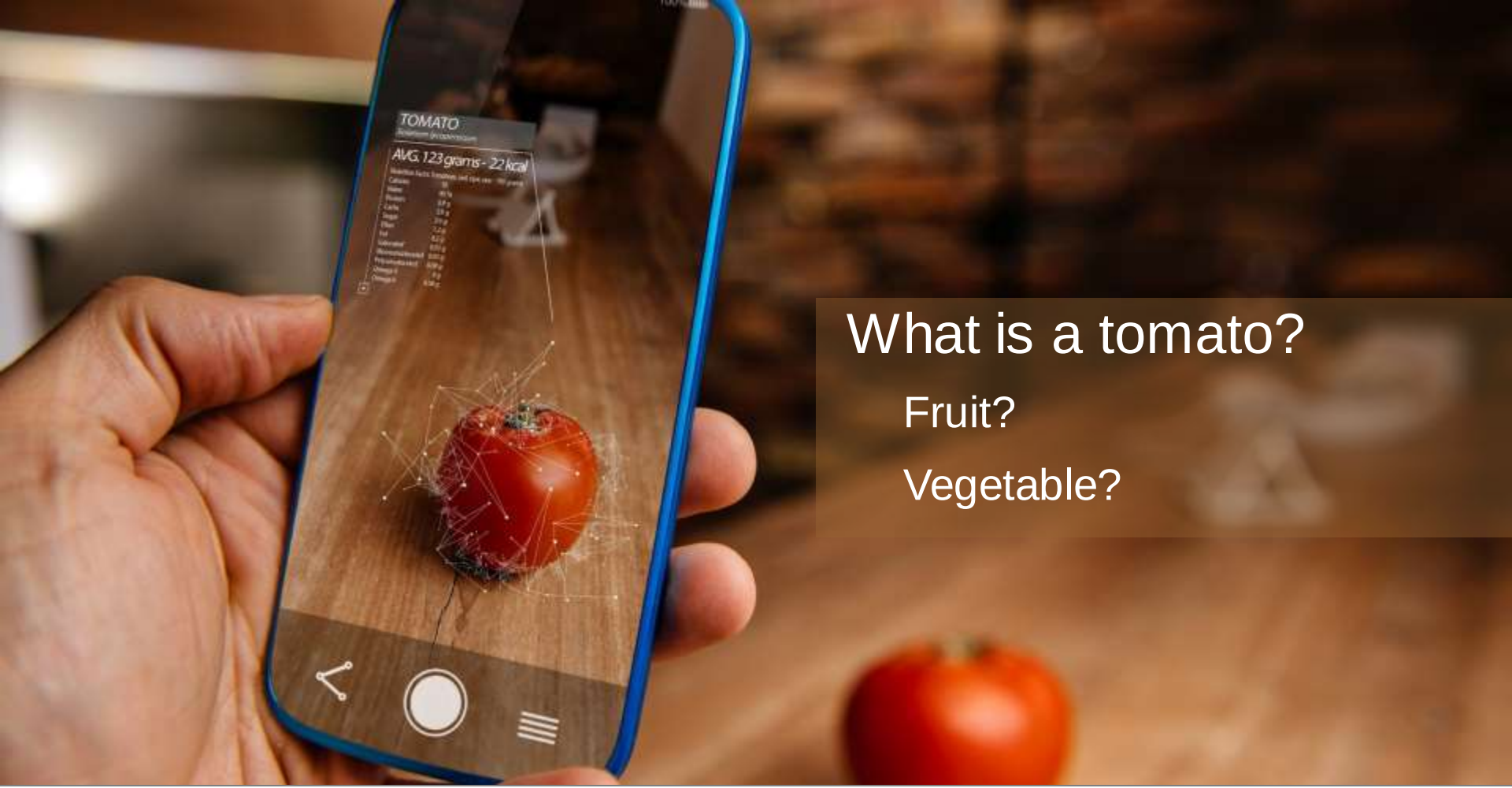
© 2023 Carnegie Mellon University

[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.

AI is data

© 2023 Carnegie Mellon University

[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.



What is a tomato?

Fruit?

Vegetable?

© 2023 Carnegie Mellon University

[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.

AI is as imperfect as the humans making it

© 2023 Carnegie Mellon University

[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.

Computer vision

Training data

Data encountered

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

© 2023 Carnegie Mellon University

[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.

Only know what taught

Training data

Unrepresentative
or incomplete training data

Data encountered

Unlikely to recognize

© 2023 Carnegie Mellon University

[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.

“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

© 2023 Carnegie Mellon University

[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.

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.

Bias in data, algorithm selection, and training.

© 2023 Carnegie Mellon University

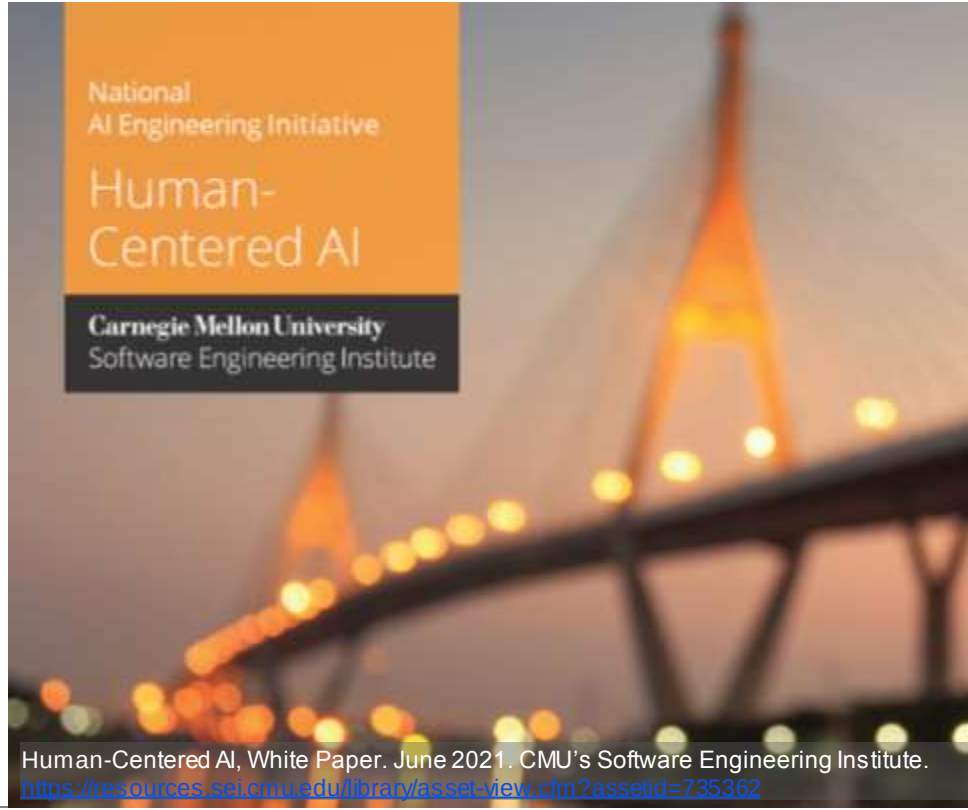
[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Design to work with, and for, people

Effective implementations

Minimize unintended consequences

1. Understand complexity of context
2. Design for human-machine teaming
3. Engage in critical oversight



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

© 2023 Carnegie Mellon University

[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.

Sense changes over time

Understand Complexity of Context

© 2023 Carnegie Mellon University

[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.

Sources of complexity

Environmental

Human

AI system capabilities

© 2023 Carnegie Mellon University

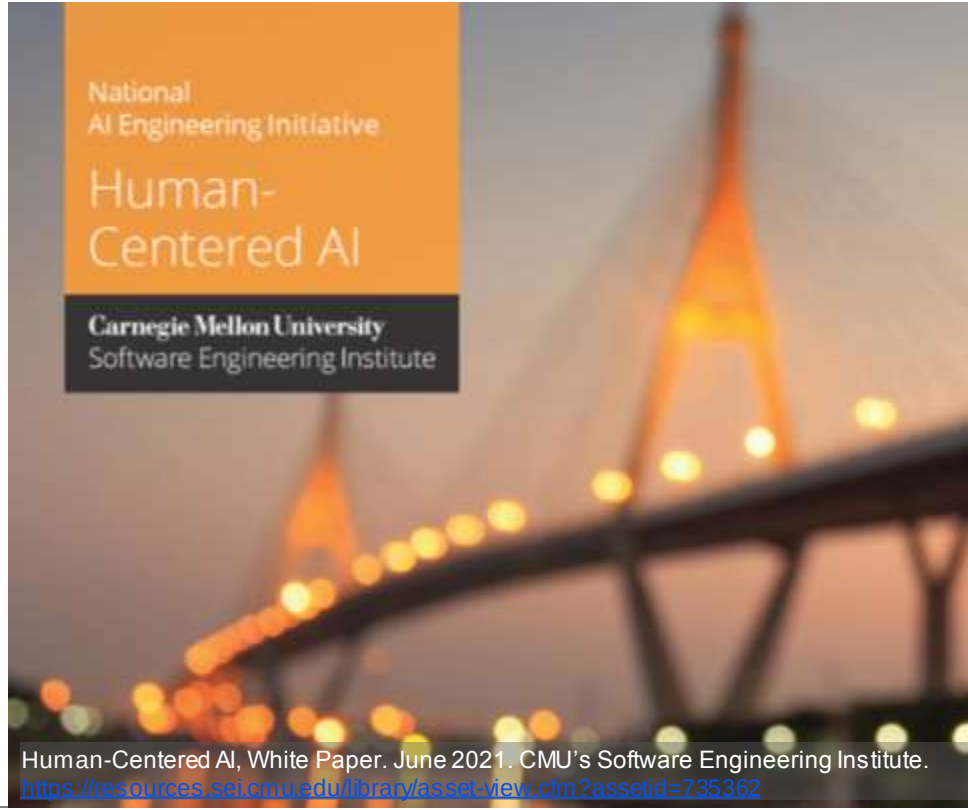
[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Research to understand context

Desired outcome, human's needs.

How do human and AI:

- learn when shifts in context have occurred?
- manage changes over time?
- adapt and evolve based on dynamic contexts?
- Share information, capabilities, and context to enable situational awareness (knowledge).



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

© 2023 Carnegie Mellon University

[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.

Speculation keeps people safe - activate curiosity

© 2023 Carnegie Mellon University

[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.

Activate Curiosity

Speculate about system misuse and abuse.

What are potential

- unintended and/or unwanted consequences?
- severe abuse and consequences?
- negative consequences for people who are in frequently marginalized groups?

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

© 2023 Carnegie Mellon University

[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.

Abusability Testing

1) Value proposition

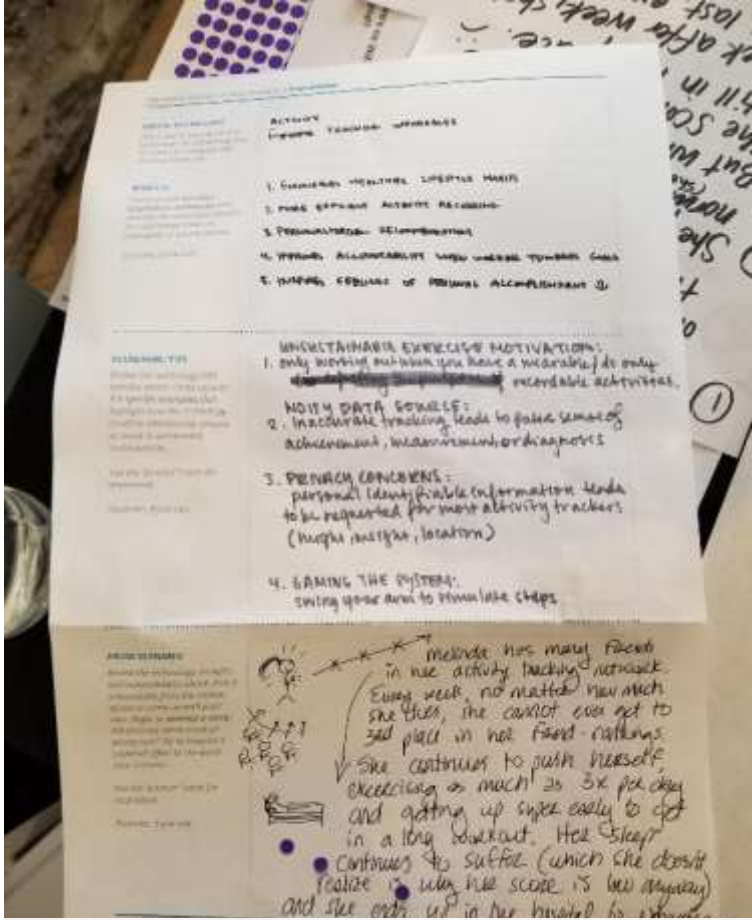
   (Potential) Benefits this tech brings to individuals or society overall

2) Vulnerabilities

   Specific examples of how tech could be misused or intentionally abused

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.

© 2023 Carnegie Mellon University

[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.

3Q-Do No Harm Framework

WHO'S NOT HERE?	HOW WILL VULNERABLE GROUPS BE NEGATIVELY IMPACTED?	WHEN THINGS DON'T WORK, HOW WILL THEY BE QUICKLY RESOLVED?
Create Inclusive Teams and Diversify Research Participants	Identify Unintended Consequences and Mitigate beforehand	Ensure the path to resolving problems is clear and fast

3Q-Do No Harm Framework, Lisa D. Dance. https://serviceease.net/3q-do-no-harm-framework

© 2023 Carnegie Mellon University

[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.

Reward team members for finding ethics bugs

© 2023 Carnegie Mellon University

[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.

What Could Go Wrong?

Card games (digital and physical)
- foster conversations around potential challenges and issues
with complex technologies (autonomous vehicles
and later AI systems).

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

© 2023 Carnegie Mellon University

[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.

Understand context

Human-centered research to identify

- Complexity and sources
- Changes over time

Inform and support designs that provide clear communication, negotiation, and coordination

© 2023 Carnegie Mellon University

[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.

Challenges

Demystifying AI to colleagues

Creating more speculative activities

Engaging teams in this hard and necessary work

Building on our experiences with UX/HCI and accessibility…

© 2023 Carnegie Mellon University

[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.

Development of tools, processes, and practices

Human-Machine Teaming

© 2023 Carnegie Mellon University

[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.

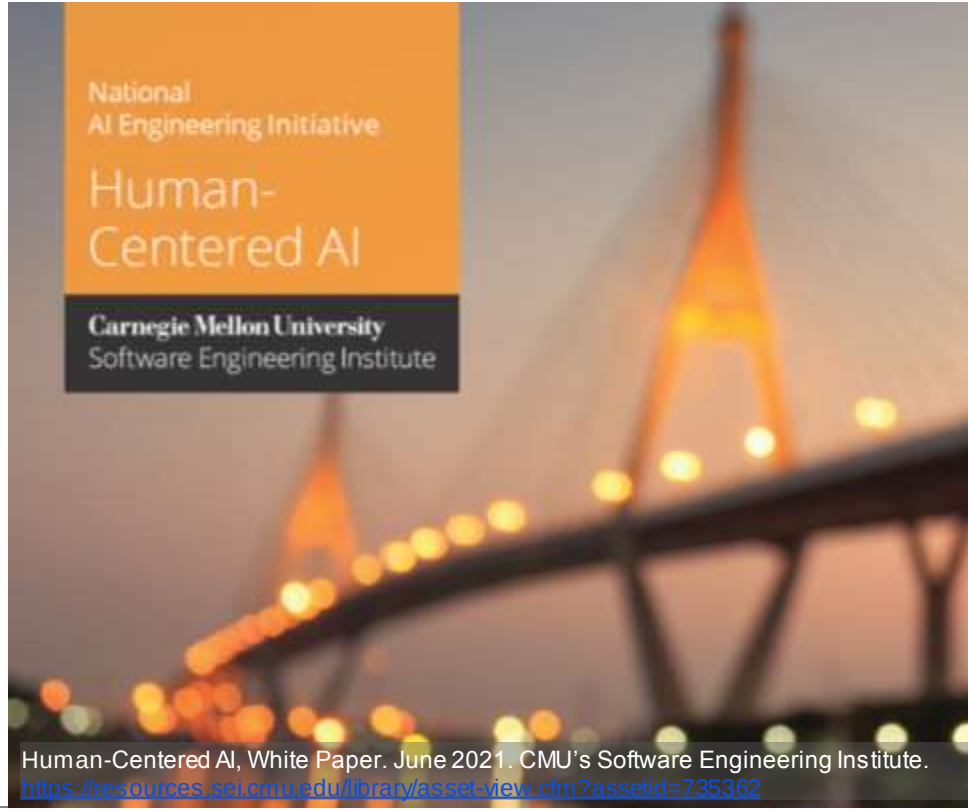
Facilitate interdependence

Human-machine teaming

Consider need for AI to:

- Augment humans' abilities
- Provide transparency regarding limitations
- Provide evidence

End-users adequately understand system and gain a *calibrated* level of trust.

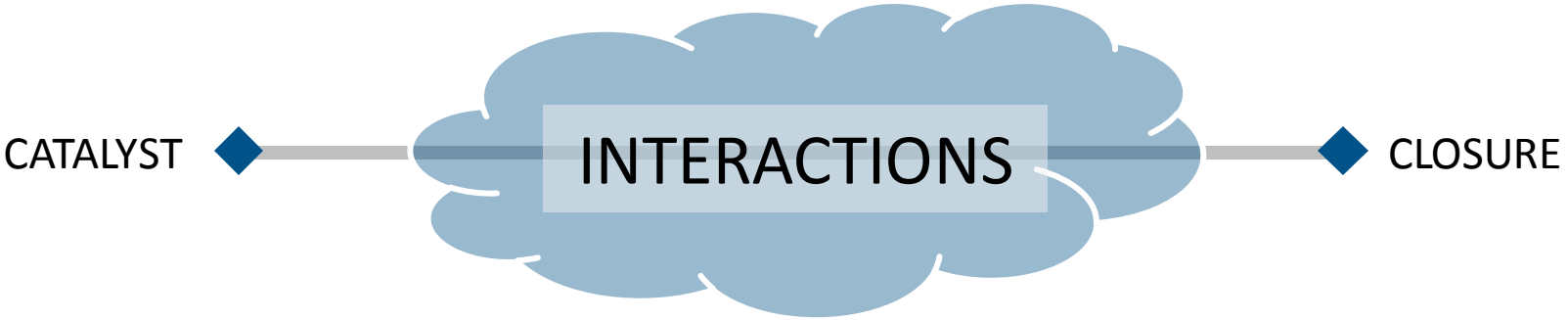


Human-Centered AI, White Paper. June 2021. CMU's Software Engineering Institute. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362

© 2023 Carnegie Mellon University

[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.

Identify collaborative activities and interactions



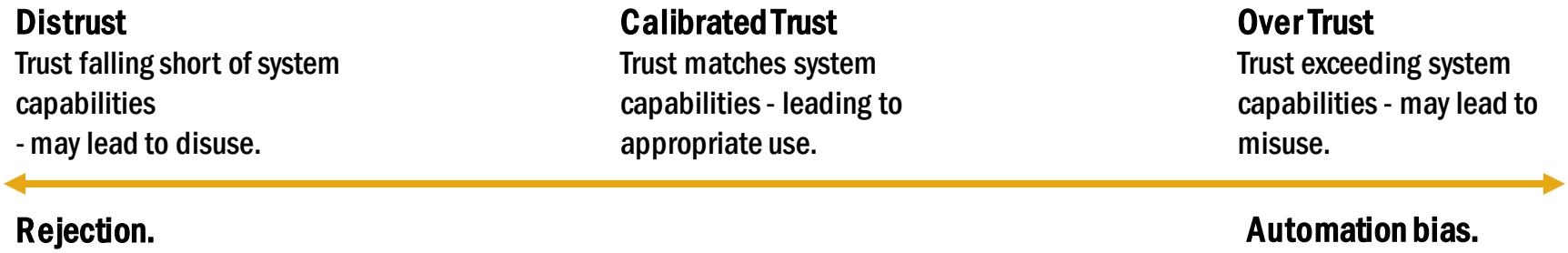
How IAs Can Shape the Future of Human-AI Collaboration
Presented on April 28-30, 2021 at the Information Architecture Conference (IAC21)

© 2023 Carnegie Mellon University

[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.

Trust is personal

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

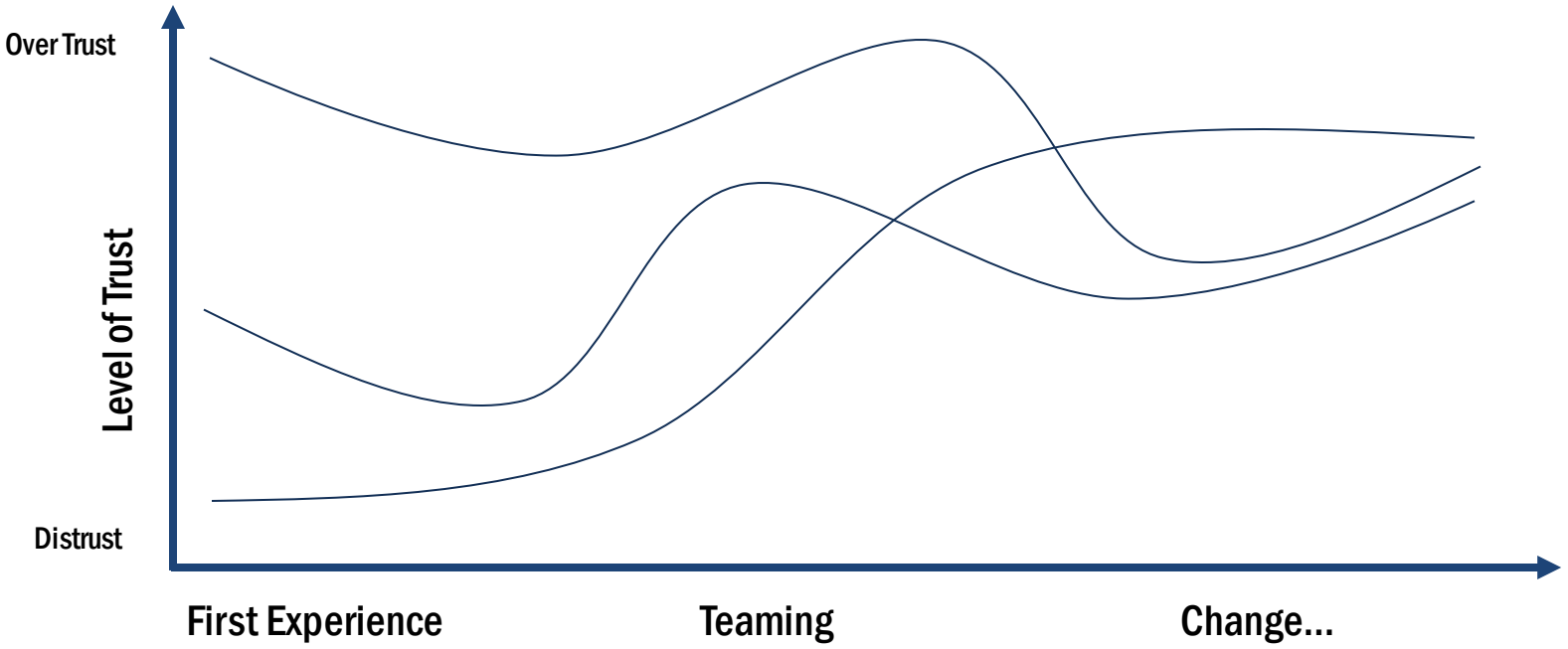


Bobbie Seppelt and John Lee. 2012. Human Factors and Ergonomics in Automation Design. In Handbook of Human Factors and Ergonomics (Fourth Edition) Chapter 59. Wiley. DOI: https://doi.org/10.1002/9781118131350.ch59

© 2023 Carnegie Mellon University

[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.

Trust changes over time



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

© 2023 Carnegie Mellon University

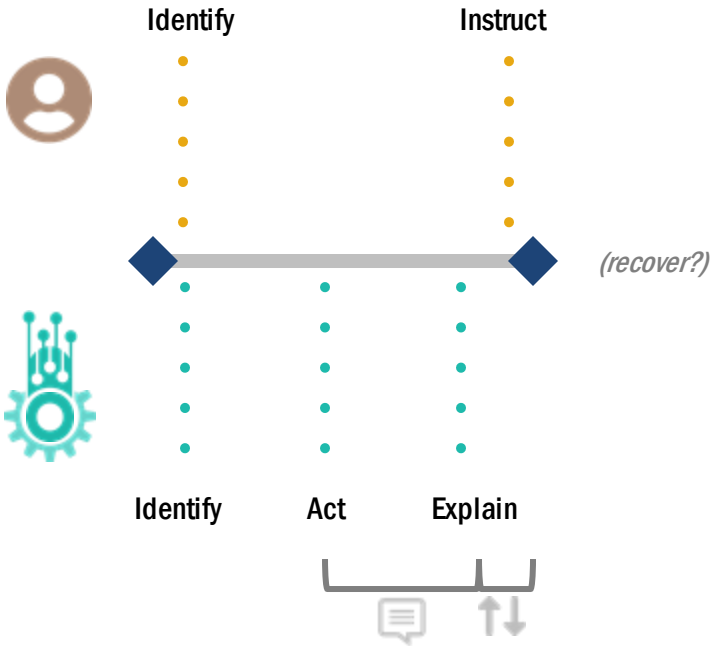
[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

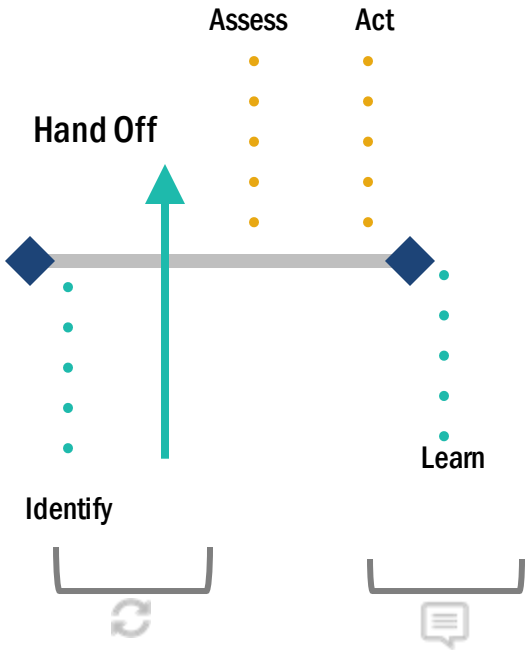
Scenario: Medical treatment decision support

© 2023 Carnegie Mellon University

[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.

Scenario: Semi-autonomous vehicle avoid road obstruction



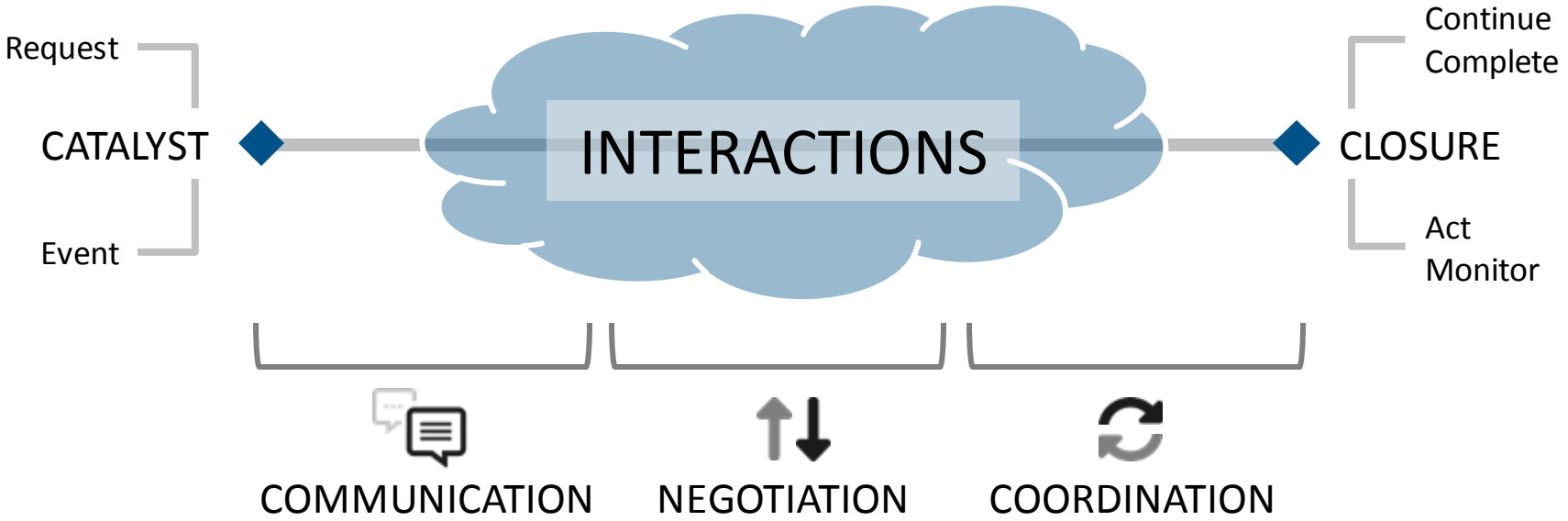


How IAs Can Shape the Future of Human-AI Collaboration
Presented on April 28-30, 2021 at the Information Architecture Conference (IAC21)

© 2023 Carnegie Mellon University

[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.

Identify collaborative activities and interactions



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

© 2023 Carnegie Mellon University

[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.

Explore range of experiences

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.

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).

© 2023 Carnegie Mellon University

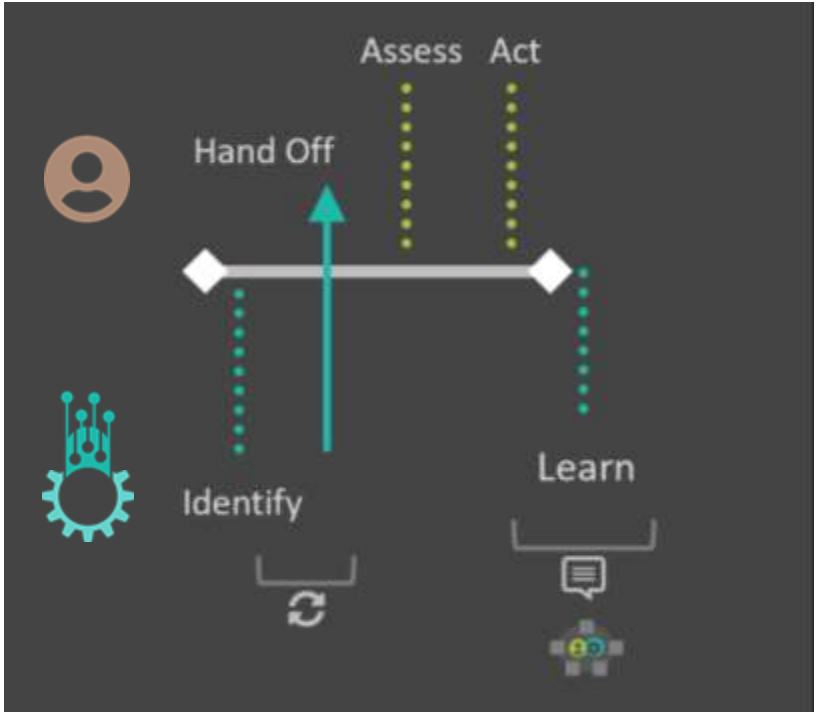
[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

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.

© 2023 Carnegie Mellon University

[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.

Identify aspects of effective team players

- How would the AI's behavior be observed?
- How would be system be easily and efficiently directed?
- What changes during busy, novel episodes?

S. W. A. Dekker and D. D. Woods. 2002. MABA-MABA or Abracadabra? Progress on Human –Automation Co-ordination. Cognition Tech Work 4, (2002) 240–244. DOI: https://doi.org/10.1007/s101110200022 Note: MABA-MABA (Men-Are-Better-At/Machines-Are-Better-At lists)

© 2023 Carnegie Mellon University

[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.

Recognize 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/

© 2023 Carnegie Mellon University

[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.

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

AI NOW INSTITUTE

Microsoft

Google

Montréal Declaration Responsible AI_

An initiative of Université de Montréal

U.S. DEPT OF DEFENSE

© 2023 Carnegie Mellon University

[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.

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

[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.

Prompt conversations

Pair technical ethics with checklists and other tools

- 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 those 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 partners 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 consent are processed 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

© 2023 Carnegie Mellon University

[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.

New uncomfortable work

"*Be uncomfortable*"

- Laura Kalbag

Ethical design is not superficial.

© 2023 Carnegie Mellon University

[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.

Design for Human-Machine Teaming

Provide transparency regarding AI limitations - boundaries and unfamiliar scenarios

Encourage appropriate trust

Speculate about misuse and abuse

Prevent or plan to mitigate situation

© 2023 Carnegie Mellon University

[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.

Challenges

Understanding what changes over time

Trust may not be measurable. Is it observable?

AI Systems are not fully able to team with humans yet, but we need to be ready!

© 2023 Carnegie Mellon University

[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.

Methods, Mechanisms, and Mindsets

Engage in Critical Oversight

© 2023 Carnegie Mellon University

[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.

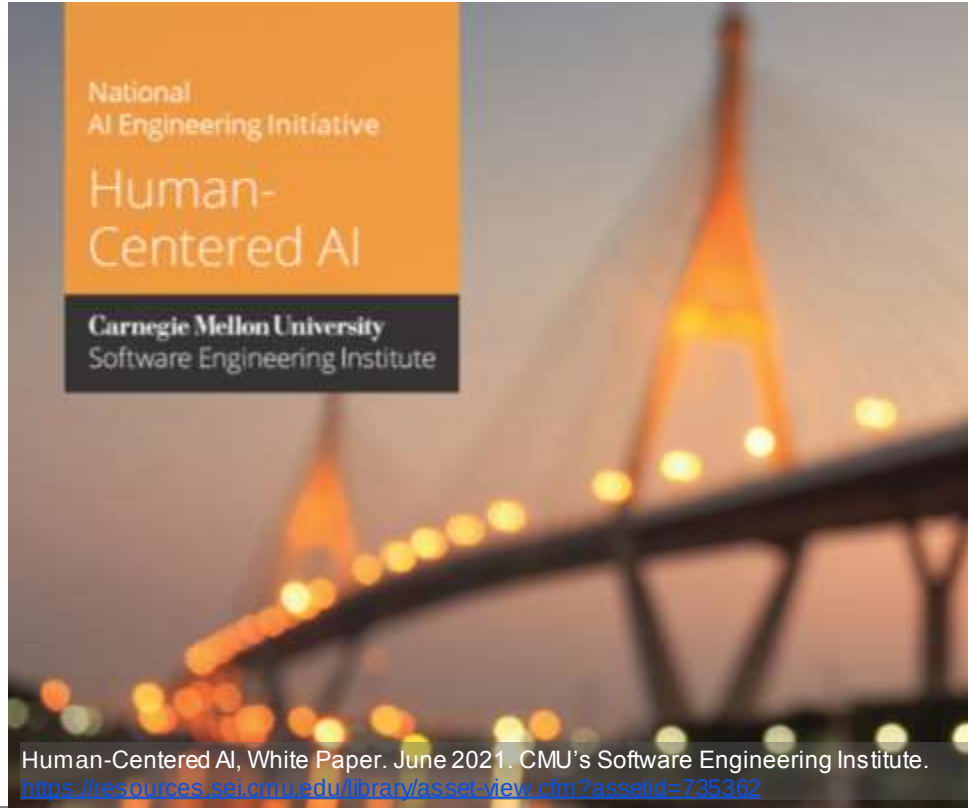
Engage in critical oversight

"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

© 2023 Carnegie Mellon University

[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.

Mitigate for bias

Understand inherent bias and amount of variance in data:

- Creator's motivation and collection process
- Data included and excluded
- Recommended uses

Bias cannot typically be removed due to our inability to identify all instances.

© 2023 Carnegie Mellon University

[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.

“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

© 2023 Carnegie Mellon University

[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.

Regular auditing to prevent harm

Dynamic systems (not stable)

Continuous human oversight required

- Probe with hypothetical cases
- Checks for bias, brittleness or potential distribution shift

© 2023 Carnegie Mellon University

[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.

Plan for Long Term Implementation and Oversight

Need continuous monitoring and evaluation for bias, brittleness, or potential distribution shift.

Nacho Kamenov & Humans in the Loop / Better Images of AI /
A trainer instructing a data annotator on how to label images / CC-BY 4.0

© 2023 Carnegie Mellon University

[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.

Leaders must establish psychological safety

© 2023 Carnegie Mellon University

[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.

Challenge: Broaden our work

Examining dynamic data and evaluating dynamic outcomes

- Is this the right data? What has changed?
- Is there evidence for calibrated trust?
- Did the system respond appropriately given the situation?
- Is the AI an effective collaborator?

We must work to define standard methods and processes for evaluating system outcomes

© 2023 Carnegie Mellon University

[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.

Responsible AI

© 2023 Carnegie Mellon University

[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.

AI has great potential, develop with caution

Future AI's *may* be trusted to substitute human cognition and abilities.

Humans must continue to be responsible for situations that involve a person's:

- Life (the use of force)
- Quality of life
- Health
- Reputation

"*AI will ensure appropriate human judgement and not replace it*" - DIB

© 2023 Carnegie Mellon University

[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.

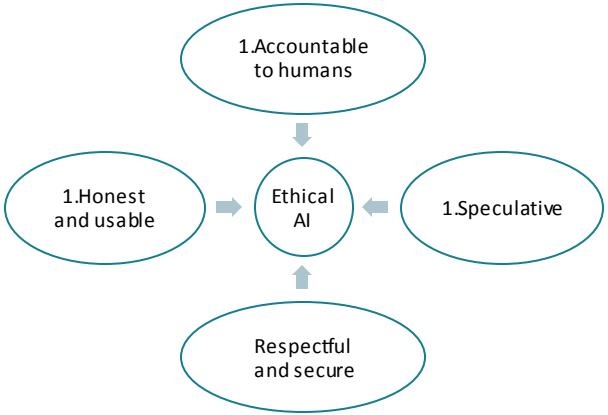
We aren’t perfect, AI won’t be perfect

Empower diverse teams, inclusive environments

Adopt technical ethics

Use responsible AI tools to encourage deep conversations

Activate curiosity; be speculative; imaginative



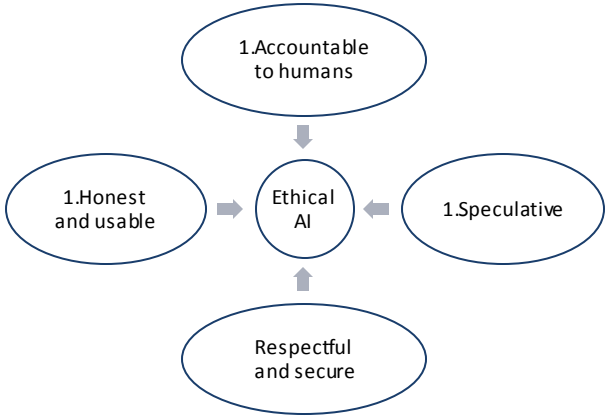
Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith.
Software Engineering Institute Blog. March 9, 2020.
Checklist and Agreement - Downloadable PDF: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620

© 2023 Carnegie Mellon University

[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.

Design to work with, and for, people

User Experience Honeycomb
Peter Morville, et al.



Responsible and Human-Centered (Ethical) AI

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

© 2023 Carnegie Mellon University

[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.

Carol J. Smith

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

AI Division
Software Engineering Institute

© 2023 Carnegie Mellon University

[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.