

# Implementing Responsible, Human-Centered AI

USER EXPERIENCE DESIGN CONSORTIUM 2023

Carol J. Smith

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



# 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.

DM23-0406

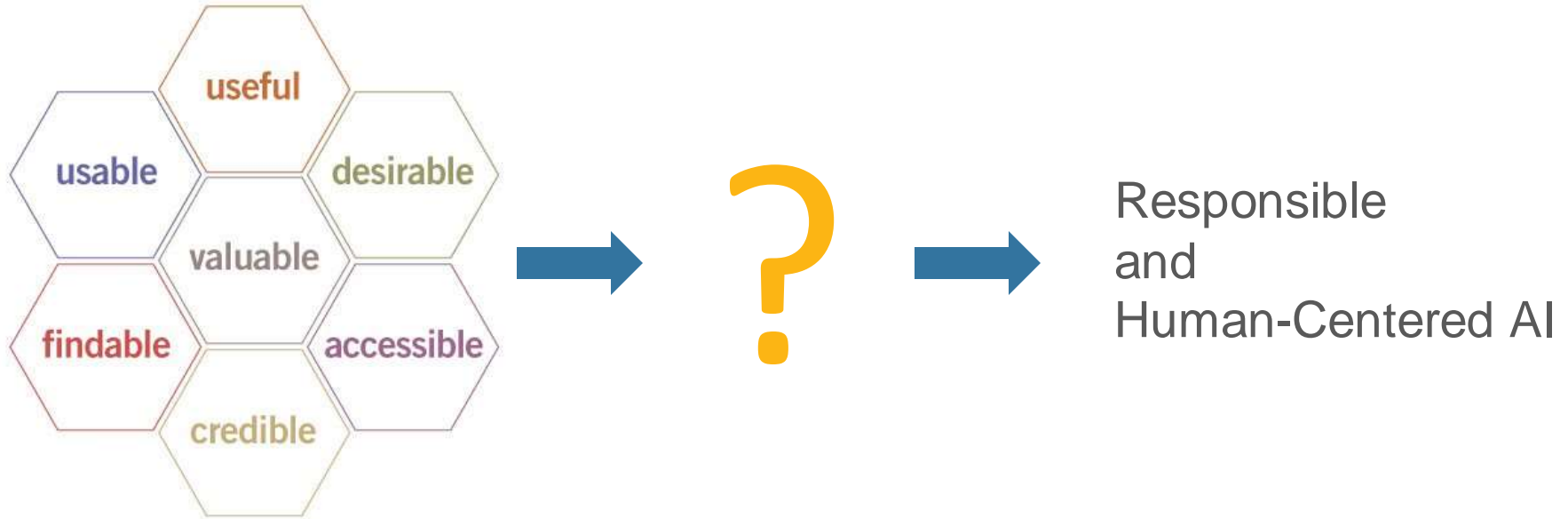
# First Machines



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



# Making Responsible and Human-Centered AI



User Experience Honeycomb  
Peter Morville, et al.

# 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?

Sensing changes over time

# Understanding Complexity of Context

# Sources of Complexity

Environmental context

Human capabilities

AI system capabilities

Research to understand

- Sources: environmental, human, and data
- Changes over time



# Understanding Context

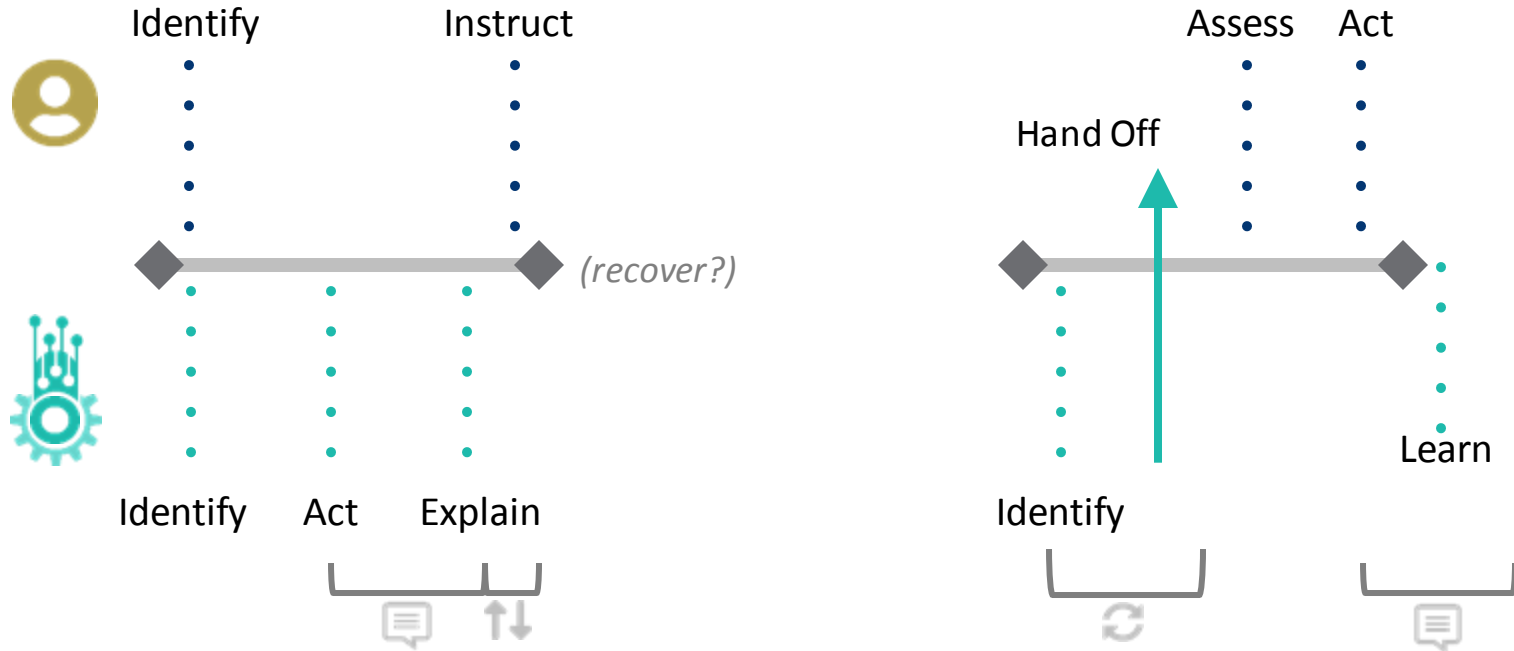
Desired outcome, human's needs  
Human and contextual factors affect  
outcome

How do human and AI:

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



# Scenario: Semi-Autonomous Vehicle Avoids Road Obstruction



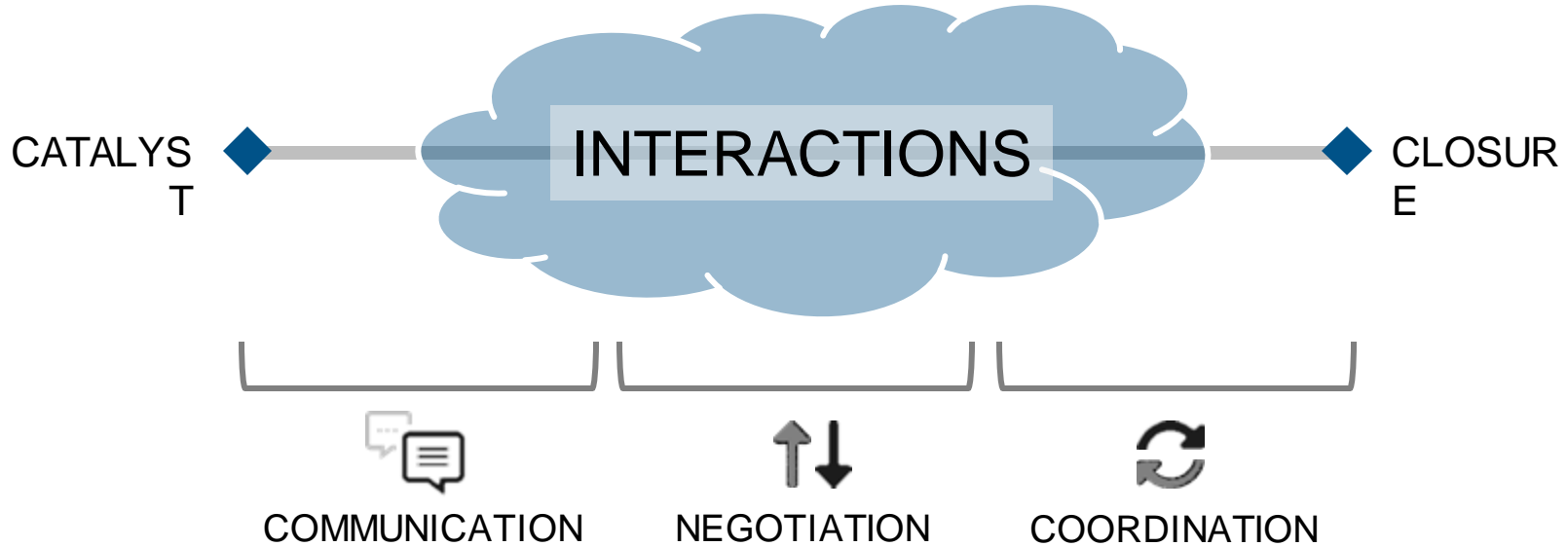
How IAs Can Shape the Future of Human-AI Collaboration  
 Presented on April 28-30, 2021 at the Information Architecture Conference (IAC21)



# Scenario: Medical Treatment Decision Support



# 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)



# Safe 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).

# Make Systems Effective Team Players

## Easy to direct

- How observable is its behavior?
- How easily and efficiently allows itself to be directed?
- Even (or especially) 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)

# 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/>

# Design for Interdependence

## Human-Machine Teaming

- people and machines
- interacting with each other.

Design AI systems to provide transparency regarding limitations.

Humans will gain *calibrated* levels of trust.



# Trust is personal

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

## **Distrust**

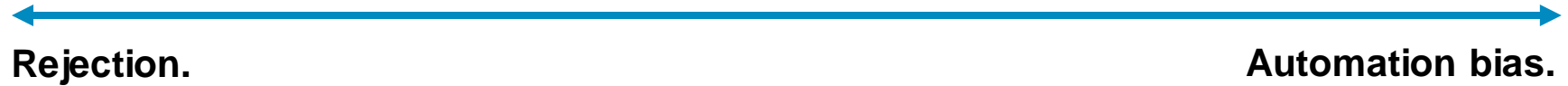
Trust falling short of system capabilities  
- may lead to disuse.

## **Calibrated Trust**

Trust matches system capabilities - leading to appropriate use.

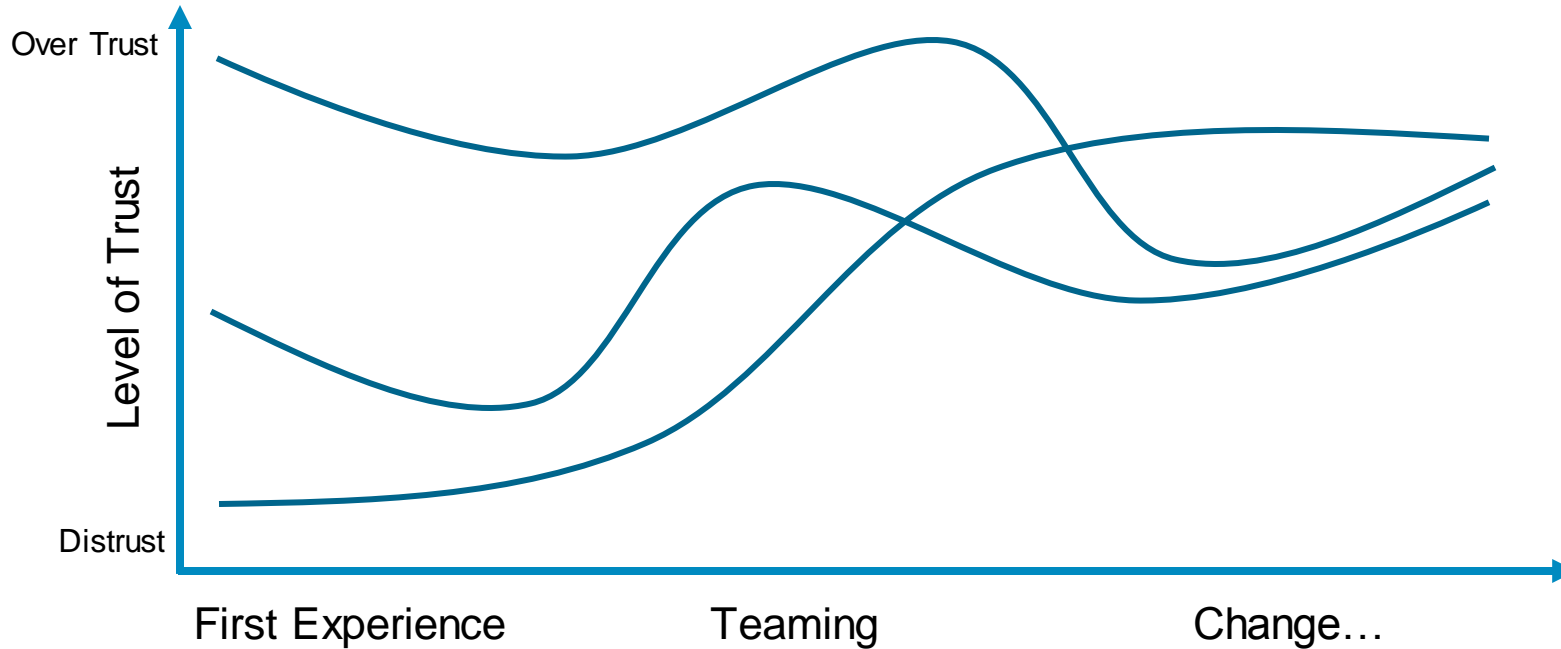
## **Over Trust**

Trust exceeding system capabilities - may lead to misuse.



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](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 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>



## Speculation keeps people safe - Activate Curiosity

# 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?\*
- How will we track our progress?
- 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](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)  
[https://unsplash.com/s/photos/business-woman-smiling?utm\\_source=unsplash&utm\\_medium=referral&utm\\_content=creditCopyText](https://unsplash.com/s/photos/business-woman-smiling?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText)



# New uncomfortable work

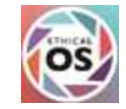
**“*Be uncomfortable*”**

- Laura Kalbag

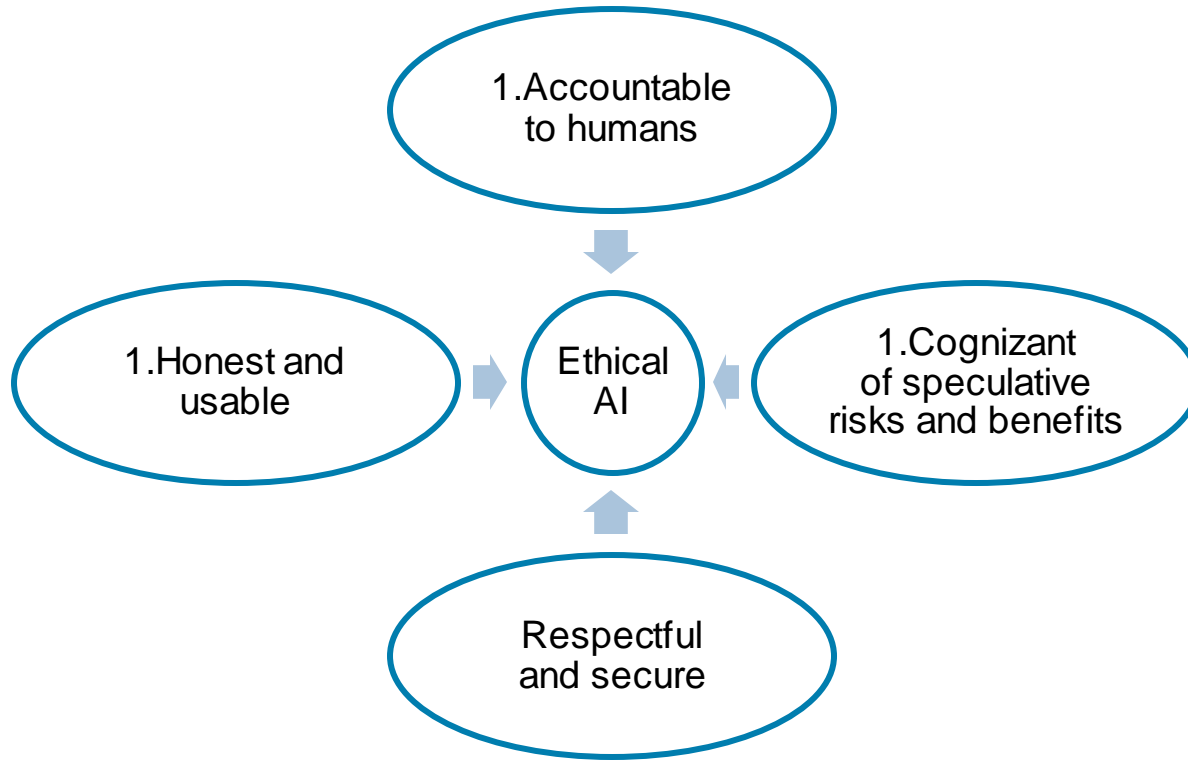
Ethical design is not superficial.

# Adopt Technology Ethics

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



# Framework for Designing Trustworthy AI



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

# Prompt conversations

## Pair checklists with technical ethics

- Bridge gaps between “do no harm” and reality
- Reduce risk and unwanted bias
- Support inspection and mitigation planning



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>

Carnegie Mellon University Software Engineering Institute

### Designing Ethical AI Experiences: Checklist and Agreement

**USE THIS DOCUMENT TO GUIDE THE DEVELOPMENT OF A TRUSTWORTHY, DE-RISKED, RESPECTFUL, HARM-FREE, AND USEFUL ARTIFICIAL INTELLIGENCE (AI) SYSTEMS WITH A DIVERSE TEAM ALIGNED ON SHARED VALUES.** 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/1910.03016>.

<p><b>We will design our AI system with the following in mind:</b></p> <ul style="list-style-type: none"> <li>Disengaged humans have the ultimate responsibility for all decisions and outcomes.             <ul style="list-style-type: none"> <li>Responsibilities are explicitly defined between the AI system and humans, and how they are shared.</li> <li>Human responsibility will be provided for those decisions that affect a person's life, quality of life, health, or reputation.</li> <li>Humans are always able to monitor, control, and deactivate systems.</li> </ul> </li> <li>Significant decisions made by the AI system will be:             <ul style="list-style-type: none"> <li>explained.</li> <li>able to be overridden.</li> <li>appealable and reversible.</li> </ul> </li> </ul>	<p><b>We work to speculatively 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 proactively research unintended consequences.</li> </ul> <p><b>We will create plans for the mitigation of the AI system, including the following:</b></p> <ul style="list-style-type: none"> <li>communication plans to share partners information with affected areas.</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>protecting privacy and data rights (only necessary data will be collected).</li> <li>providing understandable security methods.</li> <li>making the AI system robust, valid, and reliable.</li> </ul>	<p><b>We make transparency with the goal of expediting 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>Data sources have unambiguous (implied) biases, and biases are known and explicitly stated.</li> <li>Algorithms and models are open source and verifiable.</li> <li>Certification and consent are provided for humans to make decisions on:</li> <li>transparent justification for recommendations and outcomes by provided.</li> <li>strong feedback and measurable monitoring systems are provided.</li> </ul> <p><b>We value honesty and usability:</b></p> <ul style="list-style-type: none"> <li>Humans can easily discern when they are interacting with an AI system, a human.</li> <li>Humans can easily discern when and why the AI system is taking action or making decisions.</li> <li>Improvements will be made regularly to meet human needs and technical standards.</li> </ul>
--	---	---

**Team Signatures and Date**

---

**About the SEI**  
The Software Engineering Institute is a federally funded research and development organization (501)(c)(3) nonprofit organization. It is a leader in research and development in the field of software engineering and systems engineering. The SEI is the largest and oldest of the 17 national laboratories sponsored by the Department of Defense. It is a part of the University of Pittsburgh.

**Contact Us**  
1500 Locust Walk, Pittsburgh, PA 15260-1500  
 412-263-1000  
 sei@cmu.edu  
 www.sei.cmu.edu

©2020 Carnegie Mellon University. SEI-20-1108-013-1000-010

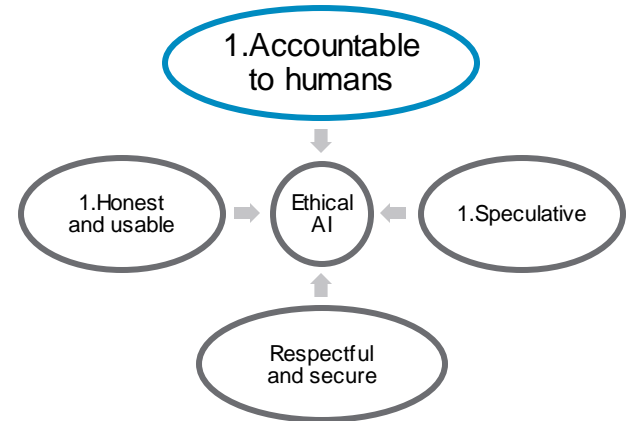
# Accountable to Humans

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

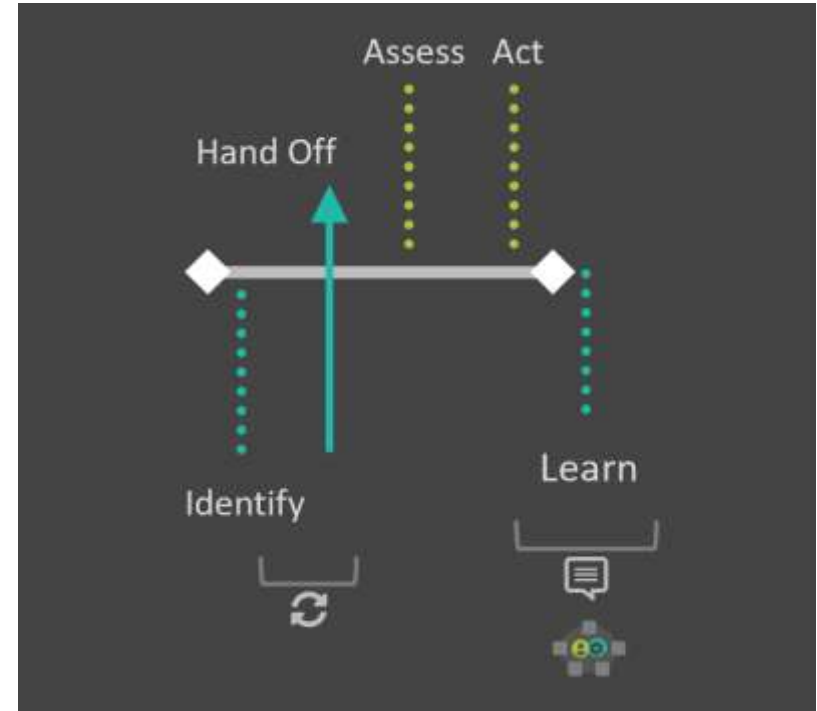
[https://www.ted.com/talks/grady\\_booch\\_don\\_t\\_rear\\_superintelligence](https://www.ted.com/talks/grady_booch_don_t_rear_superintelligence)

# Significant decisions

## Significant decisions made by system

- explained
- able to be overridden
- appealable and reversible

Responsibilities explicitly defined between people and systems.



# Cognizant of Speculative Risks and Benefits

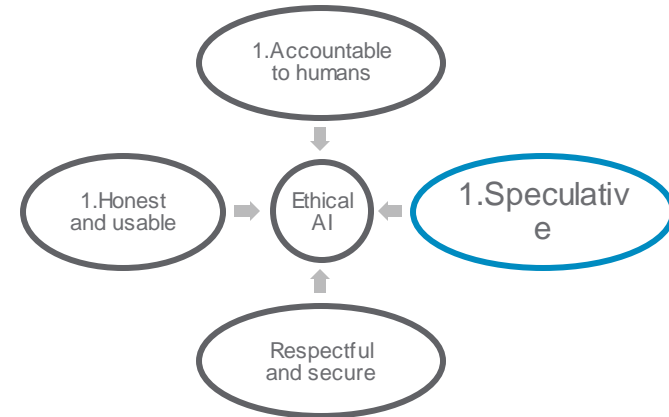
Identify full range of

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

Prevent potential harms.

Plan for unwanted consequences:

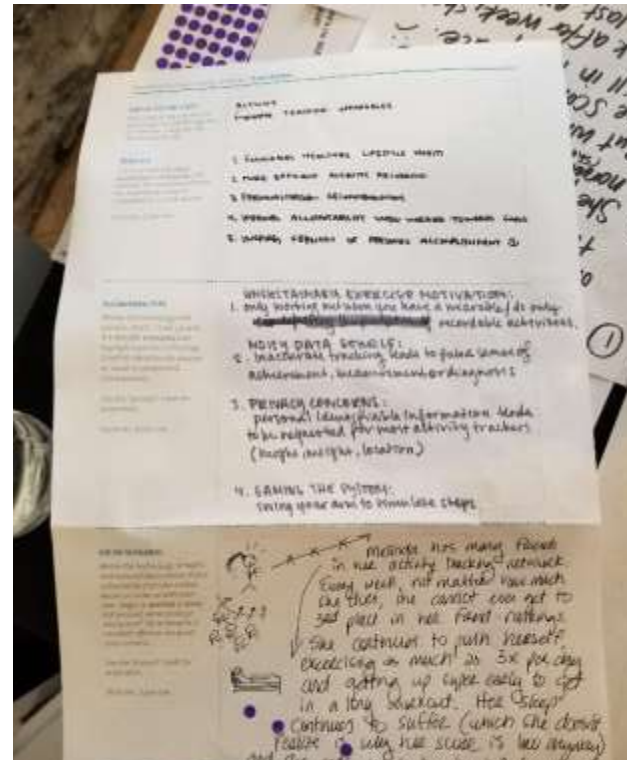
- Who can report? To whom?
- Turn off? Who notified? Consequences?



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

# Activate Curiosity

- Conduct UX research.
- Speculate about misuse and abuse.
- Potential severe abuse and consequences.
- Perspective of people in frequently marginalized groups.



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

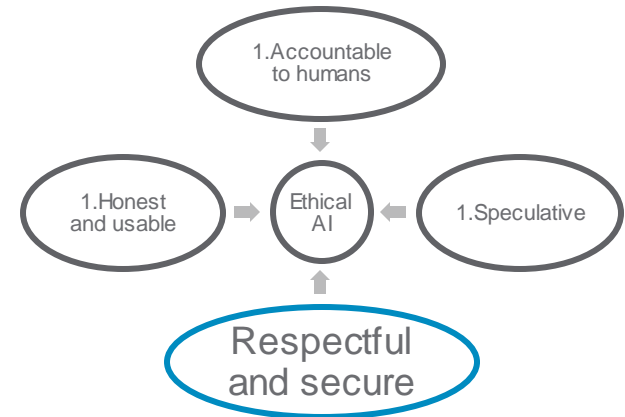
# Respectful and Secure

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

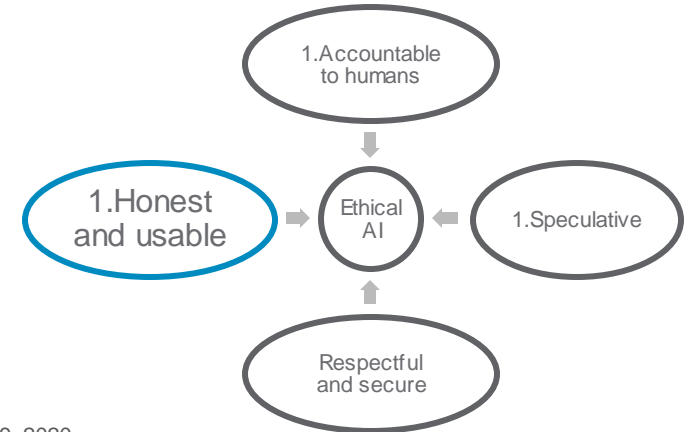
Value transparency with the goal of engendering calibrated trust.

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.

# What about authorship?



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”

## Generative models

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

## Results

- Grammatically correct and believable text
- Fabricate data (even URLs) (“Hallucinations” - misstatements)

## Controversy

- Privacy and intellectual property
- Automation bias



What is a tomato?

Fruit?

Vegetable?

# AI is as imperfect as the humans making it

# Bias in 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.

Unintended and purposeful bias

- Bias can have purpose
- Bias can be helpful

Reduce unintended/unwanted and/or harmful bias.

# Joy Buolamwini, Algorithmic Justice League

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

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



# Bias in data, algorithm selection, and training

Understand inherent bias and amount of variance.

Data:

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

Transparency and accountability.

# 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



# Defense Innovation Unit RAI Report, Guidelines, Worksheets, and Workshops



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

# Plan for Long Term Implementation and Oversight

- Training management
- Backend system support
- Continuous monitoring and evaluation
  - Bias
  - Brittleness
  - 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

Leaders  
establish  
psychological safety



# Design to work with, and for, people



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



Responsible and Human-Centered AI

User Experience Honeycomb  
Peter Morville, et al.



Carol J. Smith

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

AI DIVISION  
SOFTWARE ENGINEERING INSTITUTE

