T2: Risk, Resiliency & Responsible AI
AI World Government 2022

Implementing Responsible, Human-Centered AI

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

Twitter: @carologic @SEI_CMU_AI

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

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.

DM22-0835

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.

?

Responsible
and
Human-Centered AI

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

- Environmental, human, and information complexity
- Changes over time

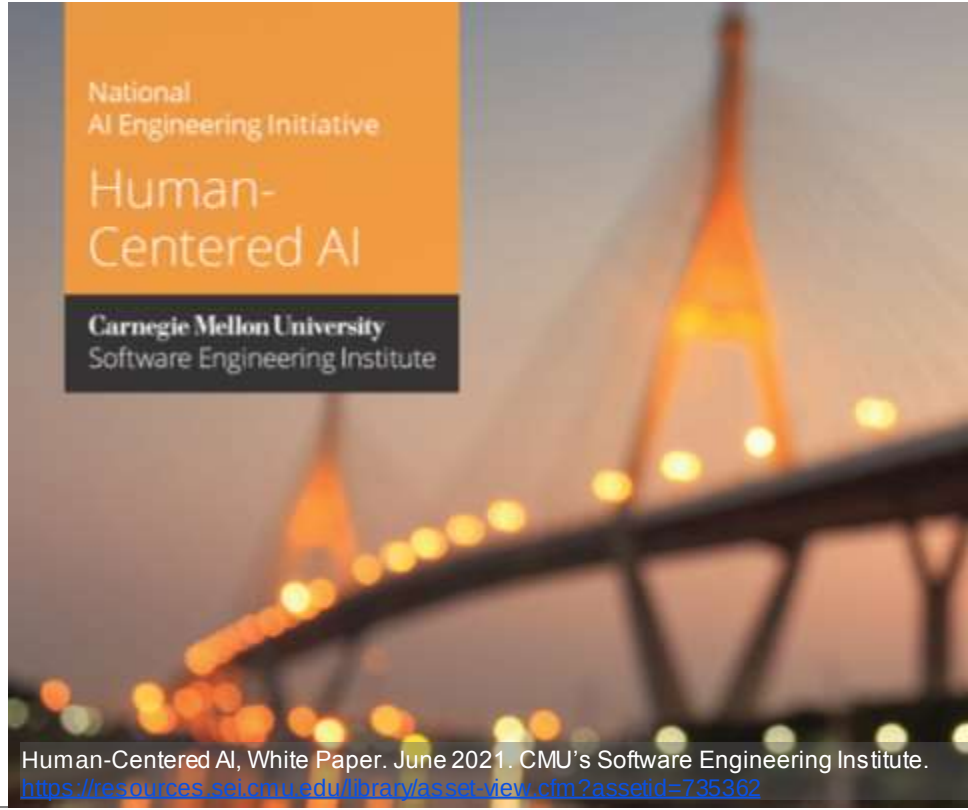
Understanding Context

Desired outcome, human's needs

Human and contextual factors affect outcome

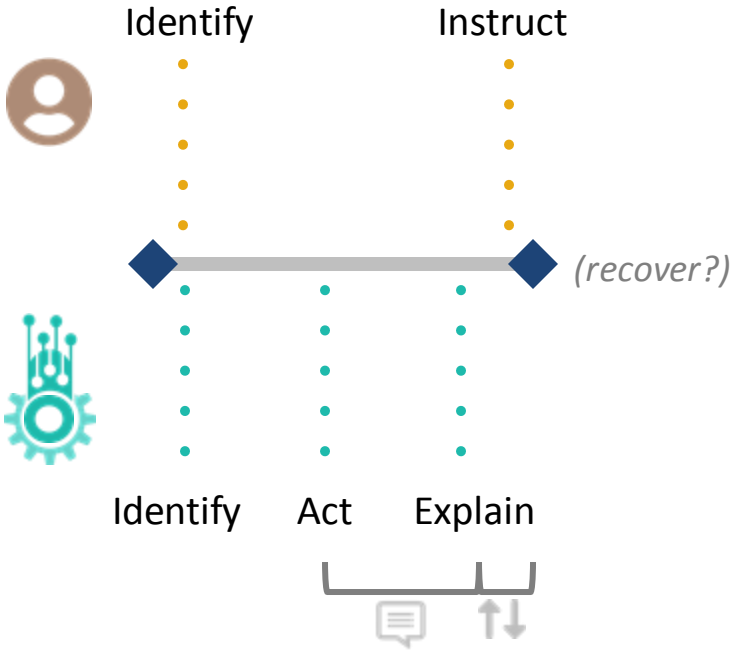
How do human and AI:

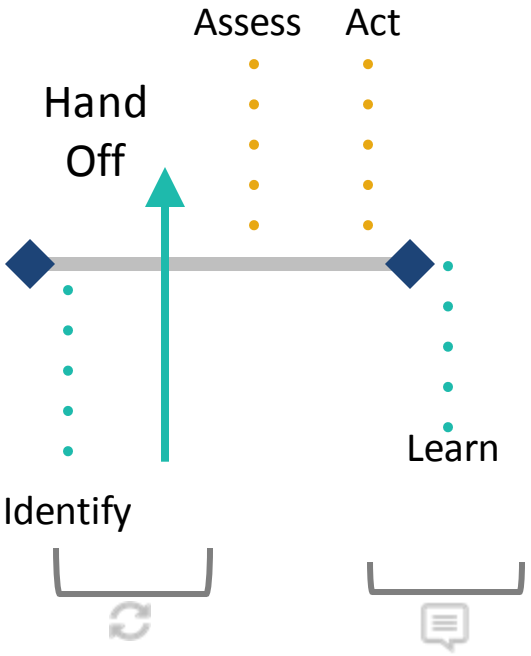
- learn when shifts in context have occurred?
- maintain clarity around operational intent?
- adapt and evolve based on dynamic contexts?



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

Scenario: Semi-Autonomous Vehicle Avoids Road Obstruction

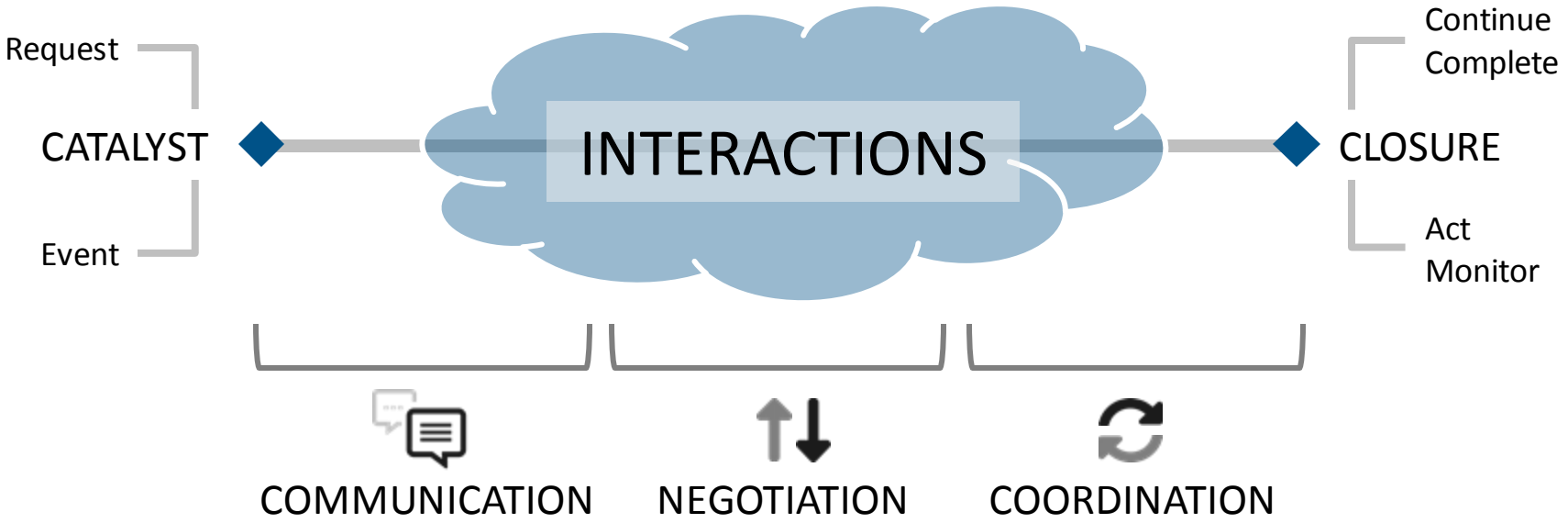




How IAs Can Shape the Future of Human-AI Collaboration
Presented on April 28-30, 2021 at the Information Architecture Conference (IAC21)

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

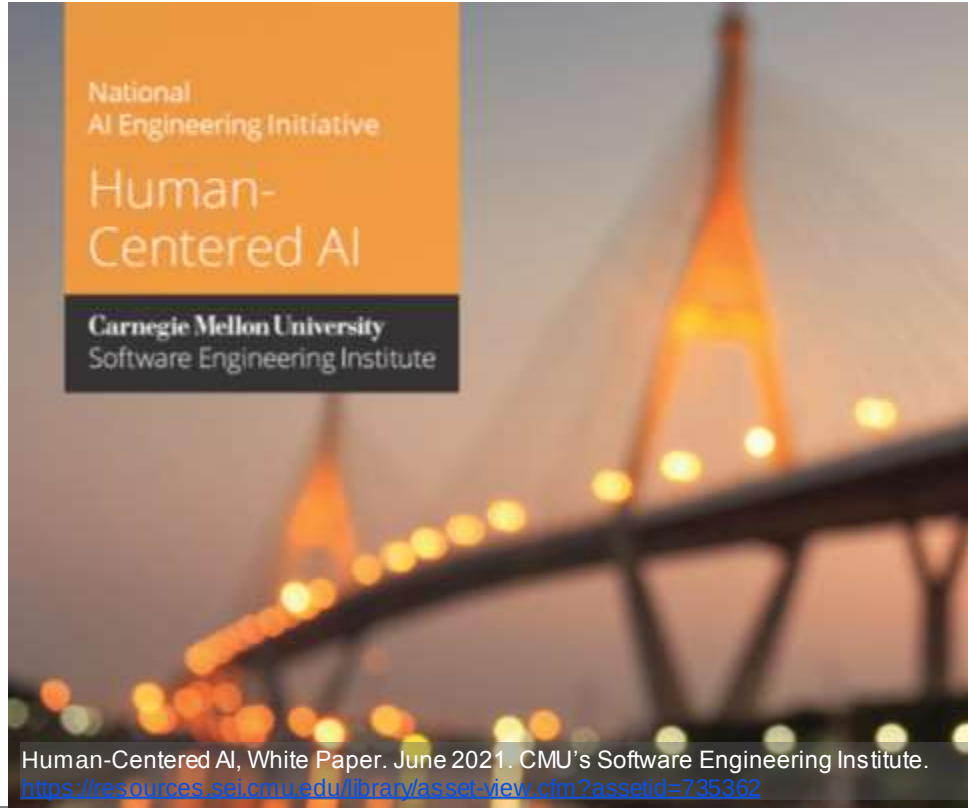
Amanda Muller and Carol Smith. 2022. Perceptions of Function Allocation between Humans and AI-Enabled Systems. UXPA 2022 (pre-print). https://uxpa2022.org/sessions/perceptions-of-function-allocation-between-humans-and-ai-enabled-systems/

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.



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

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.

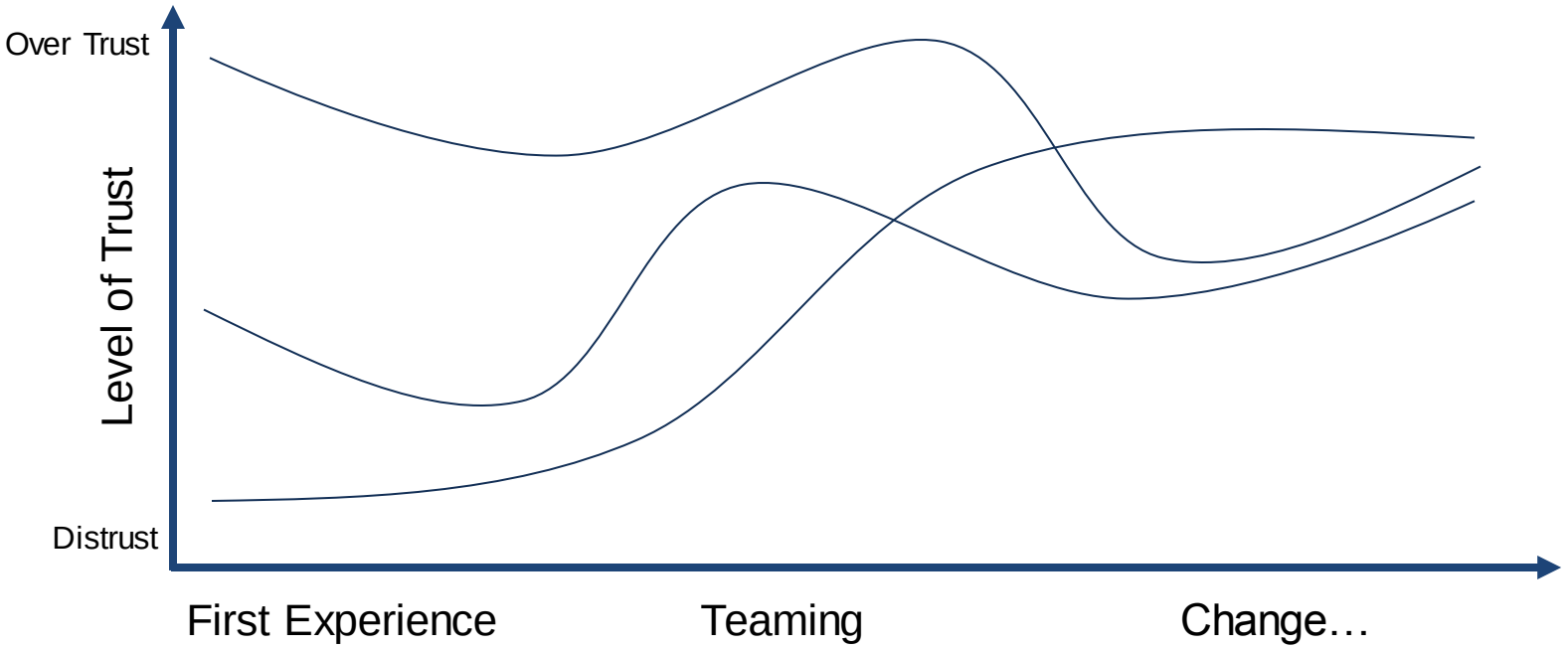
Over Trust
Trust exceeding system capabilities - may lead to misuse.

Rejection.

Automation bias.

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 On Unsplash - 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

Association for Computing Machinery

AI NOW INSTITUTE

Microsoft

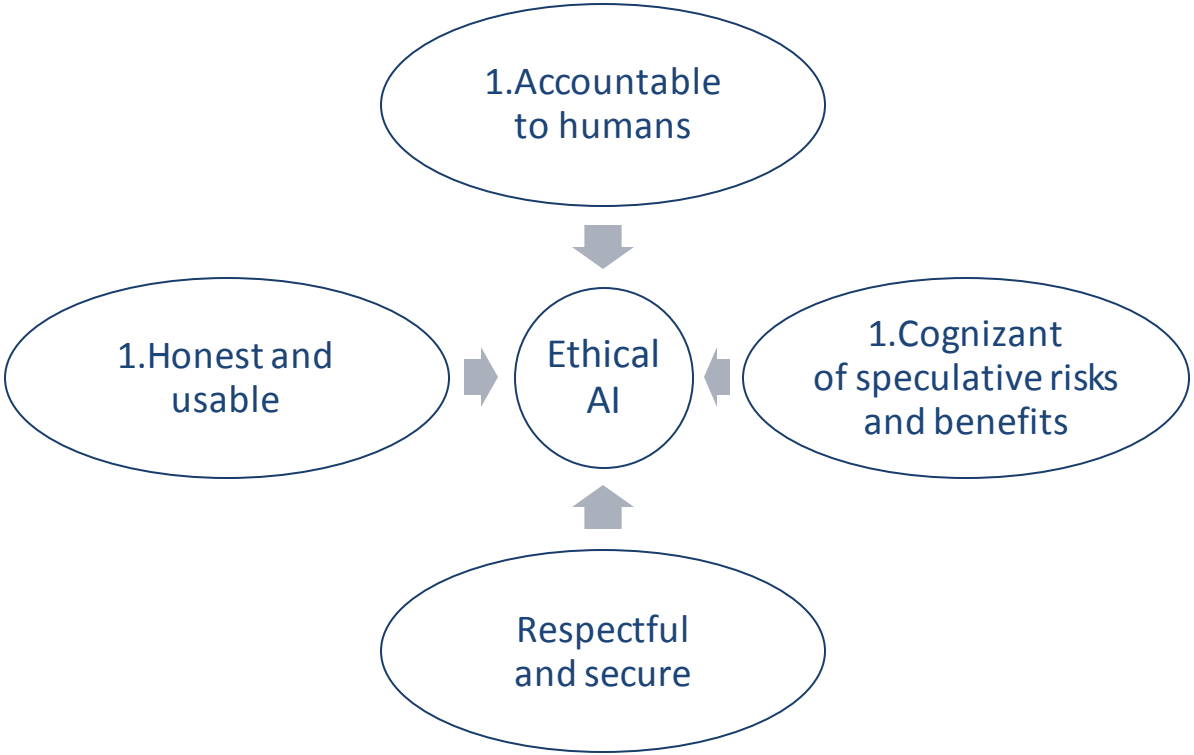
Google

Montréal Declaration Responsible AI

An initiative of Université de Montréal

U.S. DEPT OF DEFENSE

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

Carnegie Mellon University
Software Engineering Institute

Designing Ethical AI Experiences: Checklist and Agreement

USE THIS DOCUMENT TO GUIDE THE DEVELOPMENT of accountable, de-biased, 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 misunderstanding 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 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

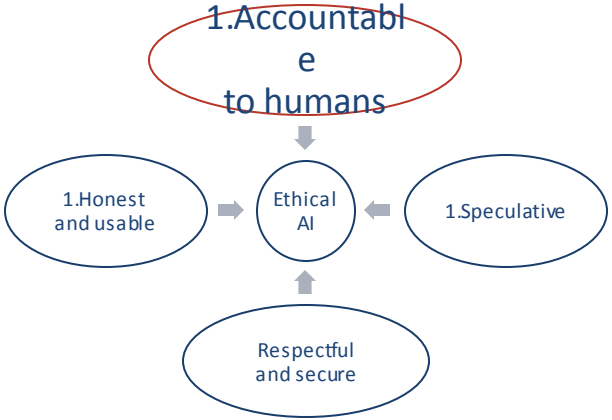
Accountable to Humans

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

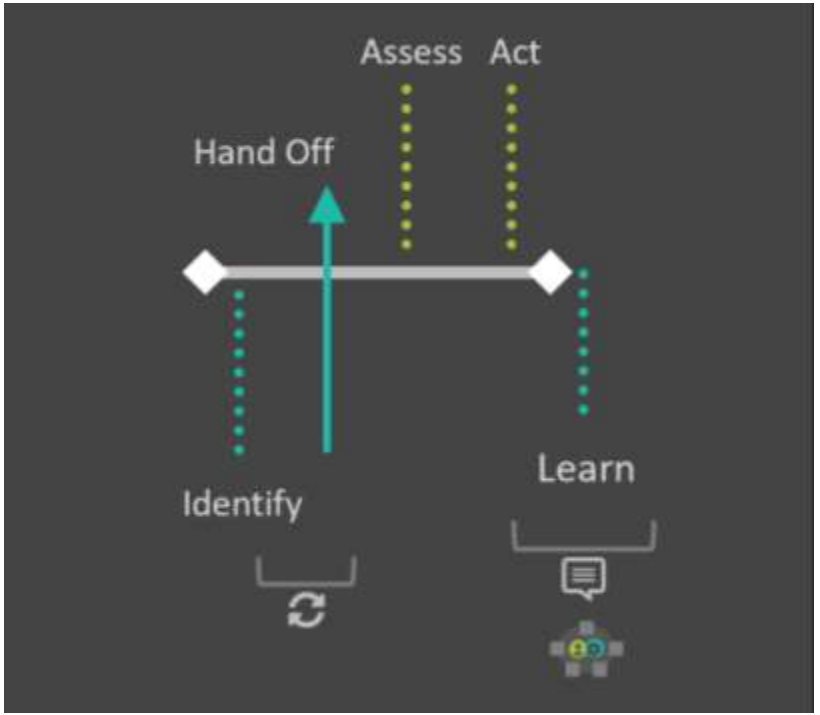
https://www.ted.com/talks/grady_booch_don_t_fear_superintelligence

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

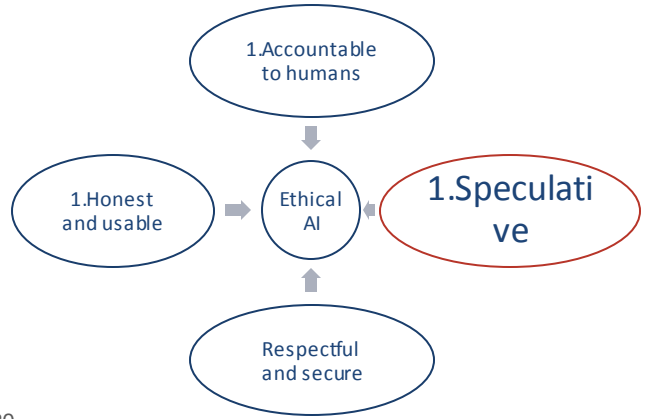
Identify full range of

- 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.

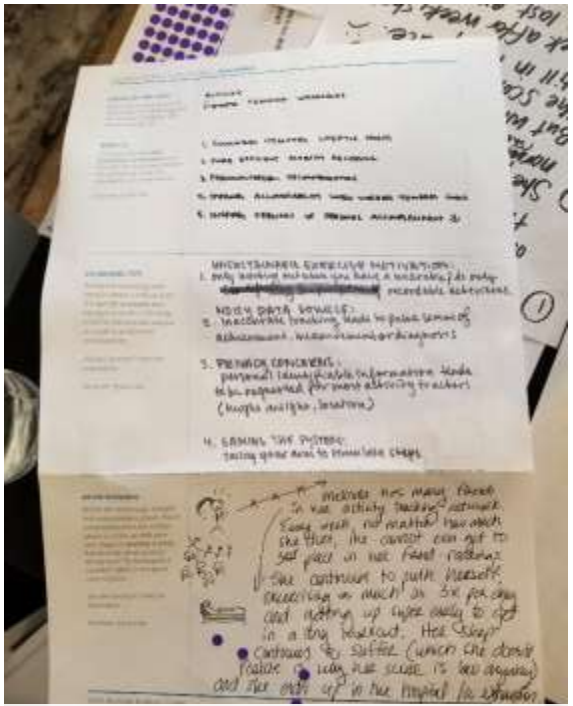
Activate Curiosity

Conduct UX research.

Speculate about misuse and abuse.

Potential severe abuse
and consequences.

Perspective of people
in frequently marginalized groups.



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

Designing Trustw orthy AI for Human-Machine Teaming. By Carol Smith. Softw are Engineering Institute Blog. March 9, 2020.

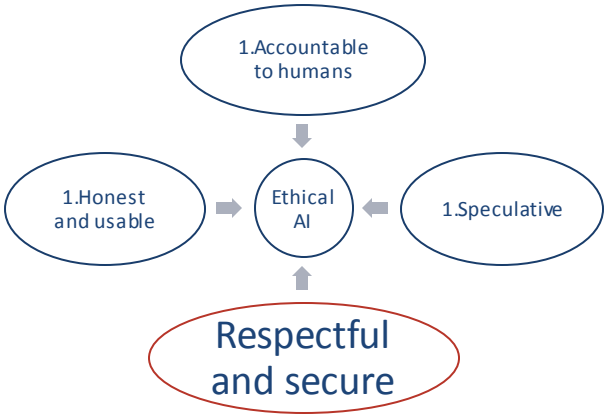
Respectful and Secure

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

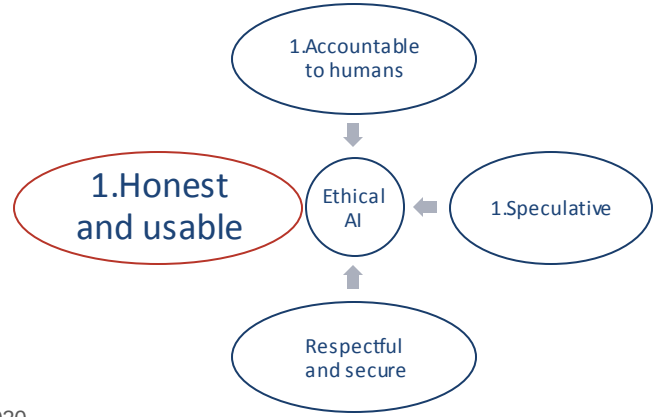
Value transparency with the goal of engendering calibrated trust.

Provide transparency regarding boundaries and unfamiliar scenarios.

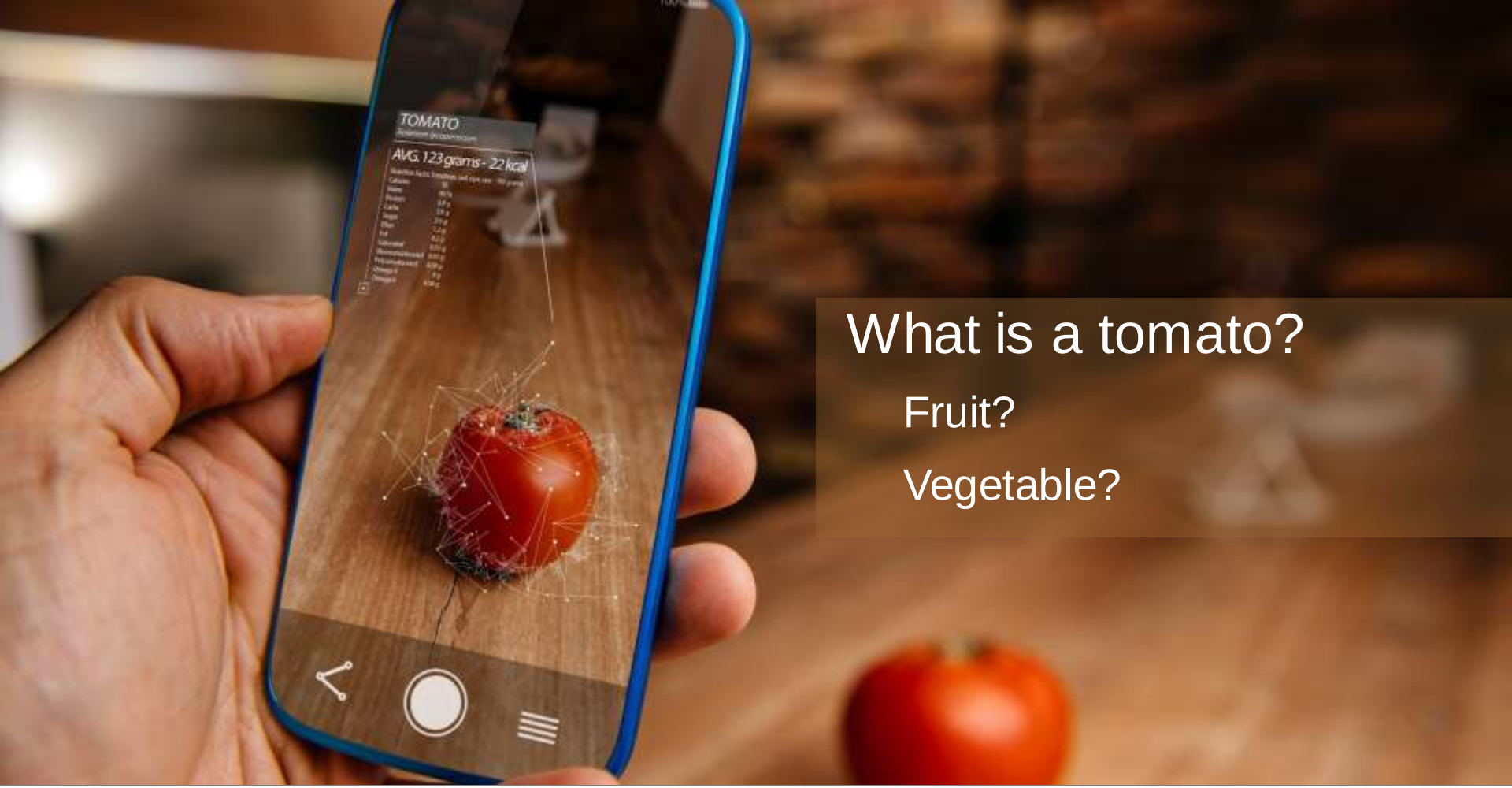
Explicitly state identity as an AI system.

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.



What is a tomato?

Fruit?

Vegetable?

AI is as imperfect as the humans making it

Bias in Image Recognition

Training data

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

Training data

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
- Bia 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

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.

Plan for Long Term Implementation and Oversight

- Training management
- Backend system support
- 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

Leaders establish psychological safety

Design to work with, and for, people

User Experience Honeycomb
Peter Morville, et al.

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

Responsible and Human-Centered AI

Carol J. Smith

Twitter: @carologic

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

CMU Software Engineering Institute,
AI Division

Twitter: @SEI_CMU_AI

Appendix

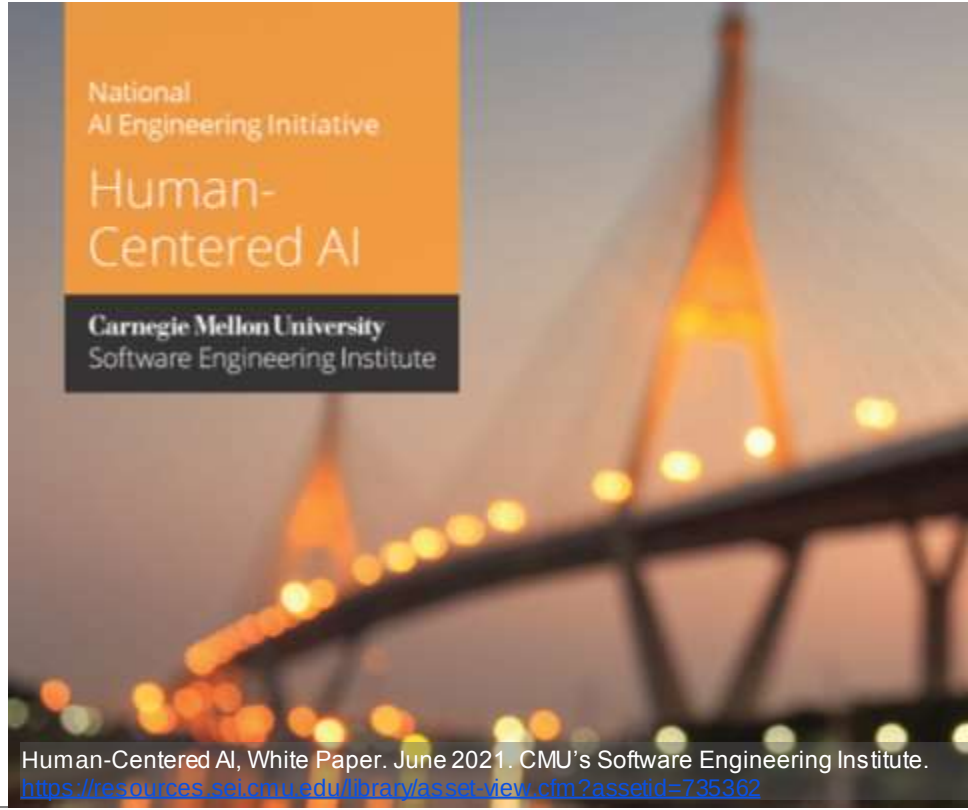
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Design to work with, and for, people

Effective implementations

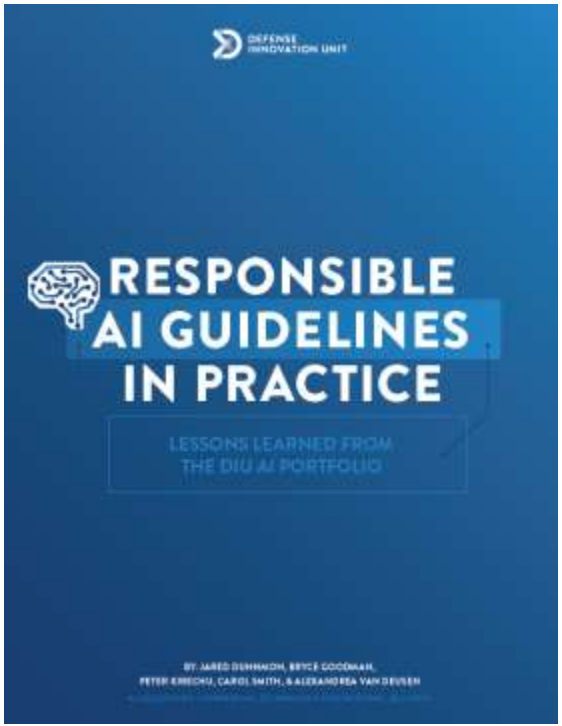
Minimize unintended consequences

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

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