Analytics Exchange: HCAI and HCI
Northwestern University, May 20, 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.

DM22-0439

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?

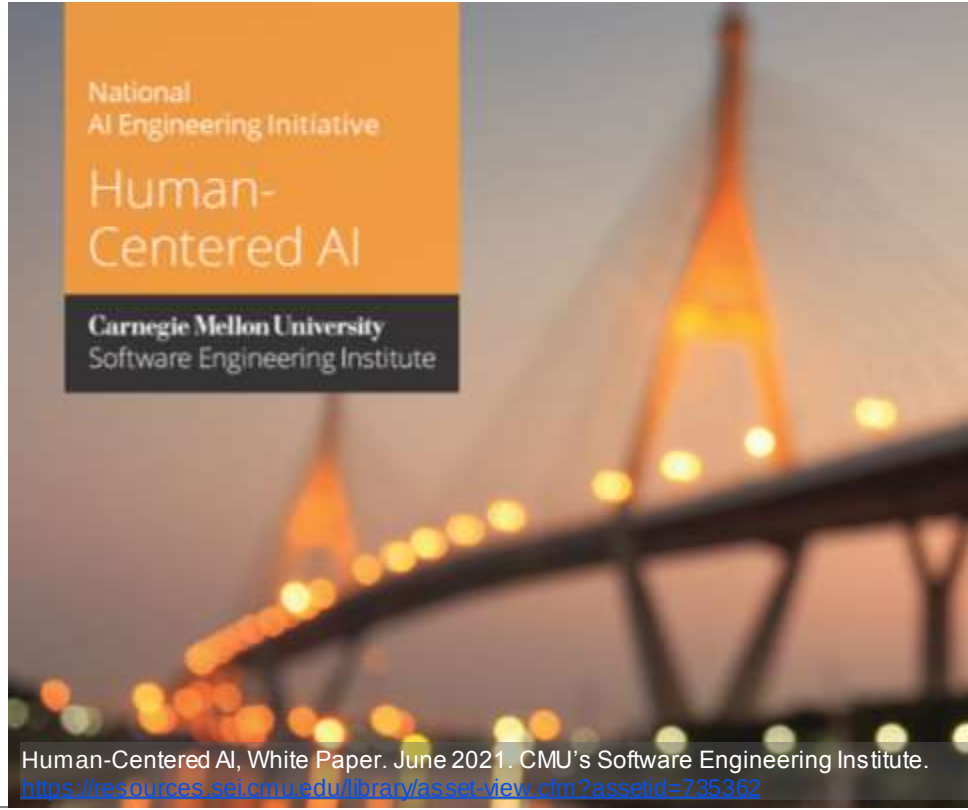
How will we know we've made improvements?

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

Sensing changes over time

Understanding Complexity of Context

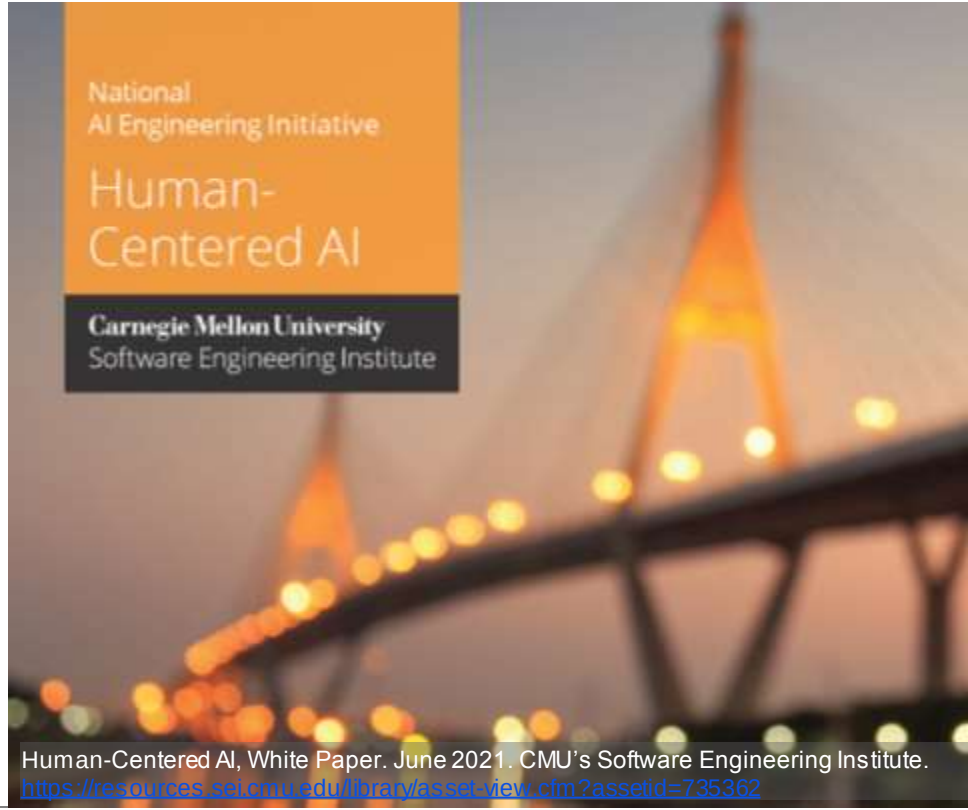
Understanding Complexity of Context

Desired outcome, human's needs

Human and contextual factors affect outcome

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

Complexity

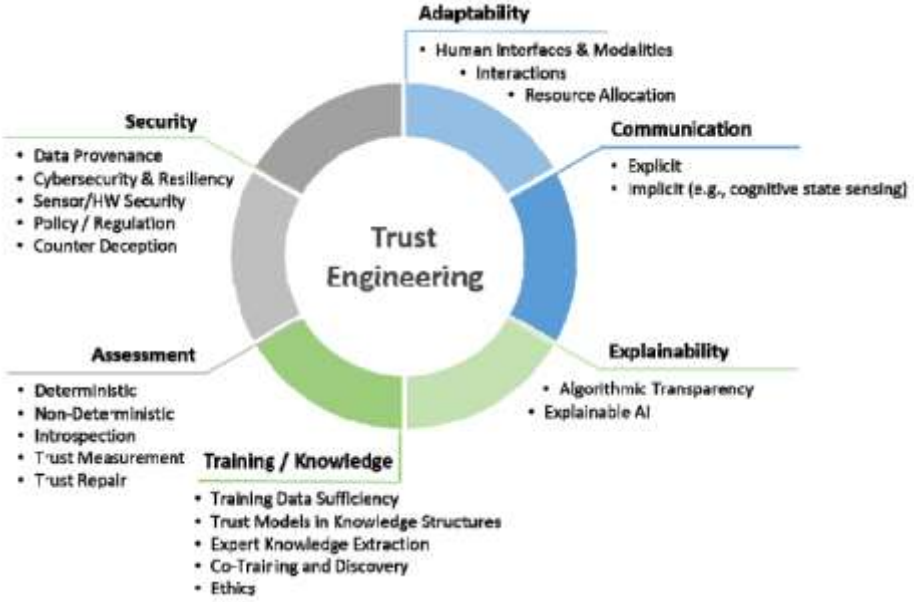
Environmental context

Human context

AI system capabilities

Information

Trust Engineering for Human-AI Teams



Design components

- Security
- Adaptability
- Communication
- Explainability
- Training/Knowledge
- Assessment

Neta Ezer, Sylvain Bruni, Yang Cai, Sam J. Hepenstal, Christopher A. Miller, and Dylan D. Schmorrow. 2019. Trust Engineering for Human-AI Teams. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 63, no. 1 (November 2019): 322–26. https://doi.org/10.1177/1071181319631264.

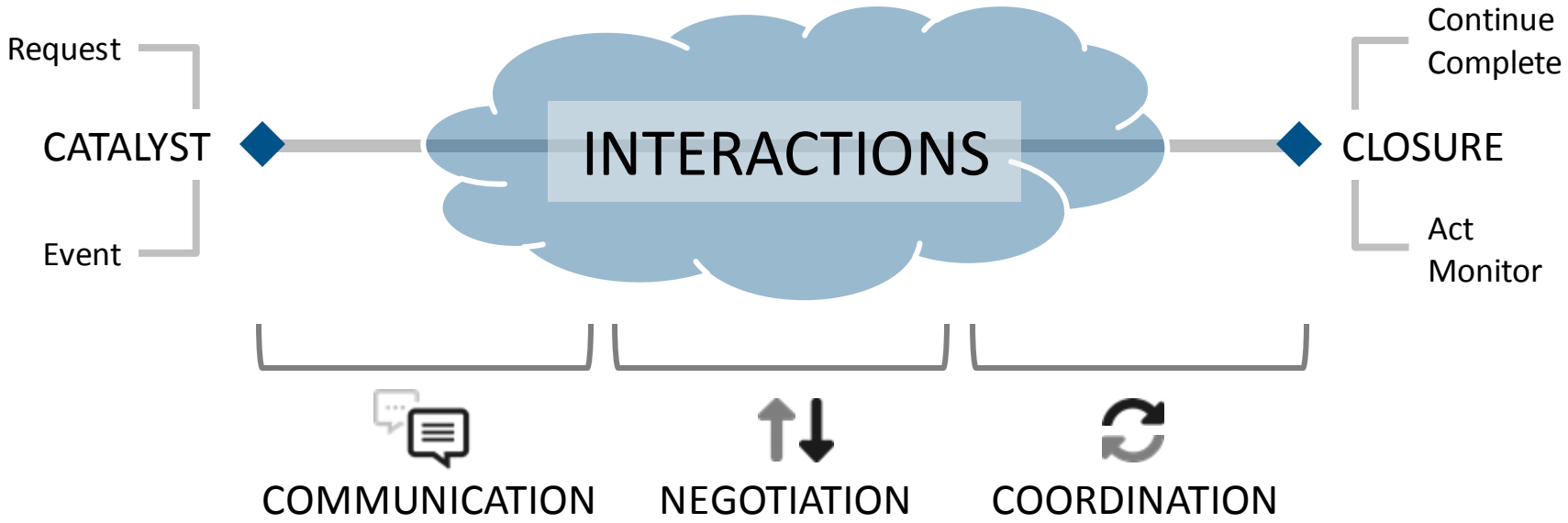
Collaborative Activities and Interactions

Length of interactions
- Short and hectic
- Longer, cyclical - iterative

Collaboration requires clear
- Communication
- Negotiation
- Coordination

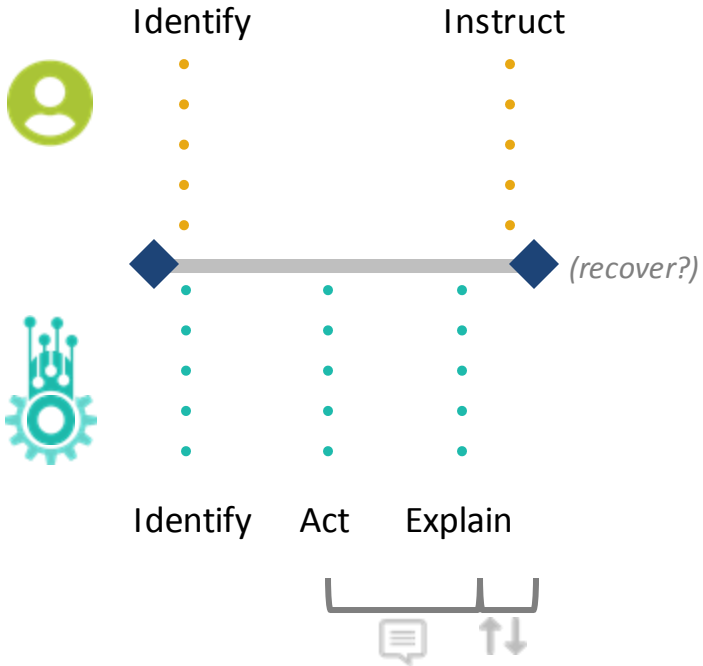
How IAs Can Shape the Future of Human-AI Collaboration
Presented on April 28-30, 2021 at the Information Architecture Conference (IAC21)

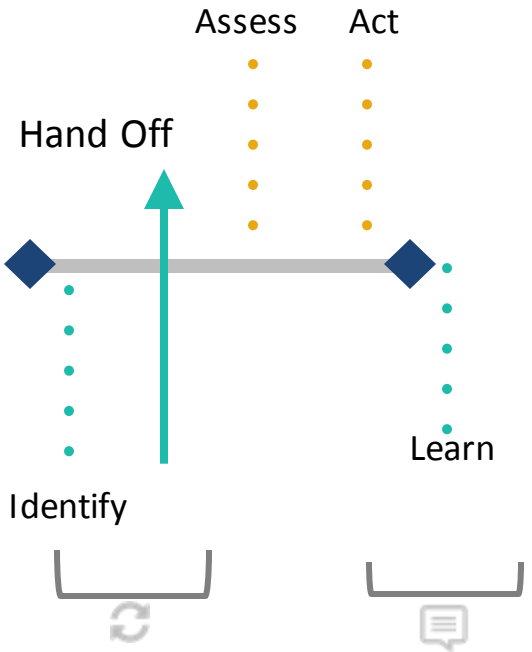
Time Cycles



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 Avoiding Road Obstruction





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

Decision making for medical treatment

Potential factors

- How much information is already known?
- Stage of disease?
- Specifics of the patient's health, stage of disease, family situation, insurance status, etc.?
- How is new information integrated and how does that change interactions?

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.

Don’t rely on operators to detect errors
and recover before an accident – it isn't realistic.

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 better at:

- 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/

Context

Human-centered research to

- Understand complexity (environmental, human and information)
- Changes over time

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

Challenges

Demystifying AI to colleagues

Getting time and budget for UX research

Using existing use patterns wisely

Understanding what changes with regard to time cycles

Development of tools, processes, and practices

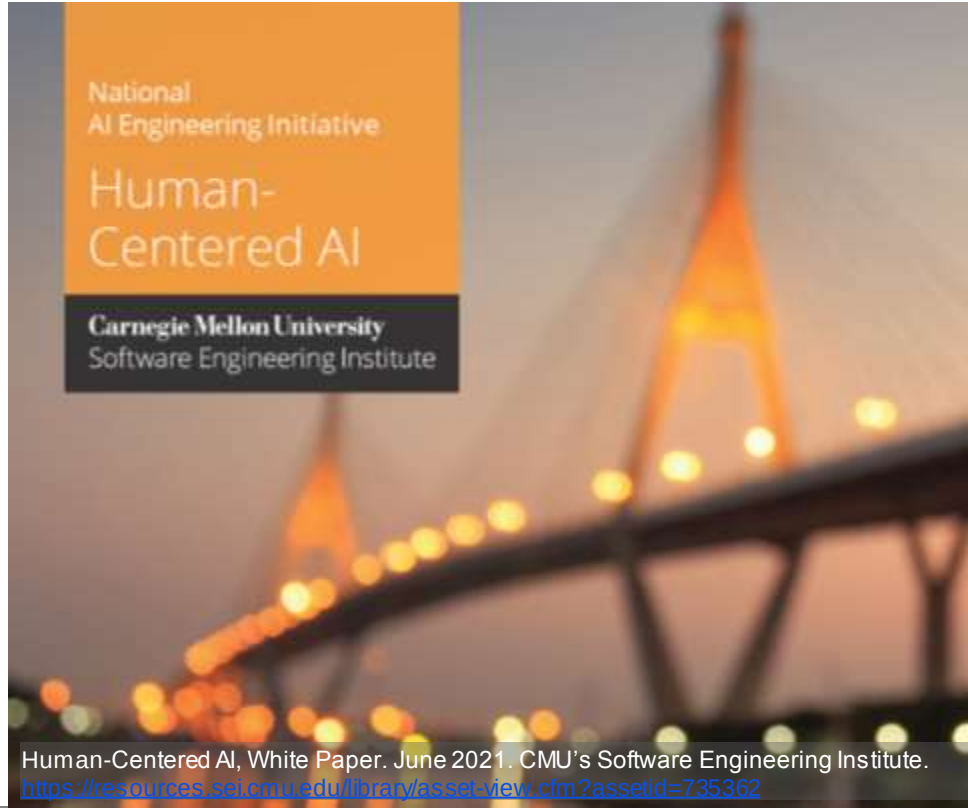
Design for Human-Machine Teaming

Design for HMT - Interdependence

People interacting with and understanding systems

Gaining *calibrated* levels of trust

Design AI system to provide transparency regarding AI limitations



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.

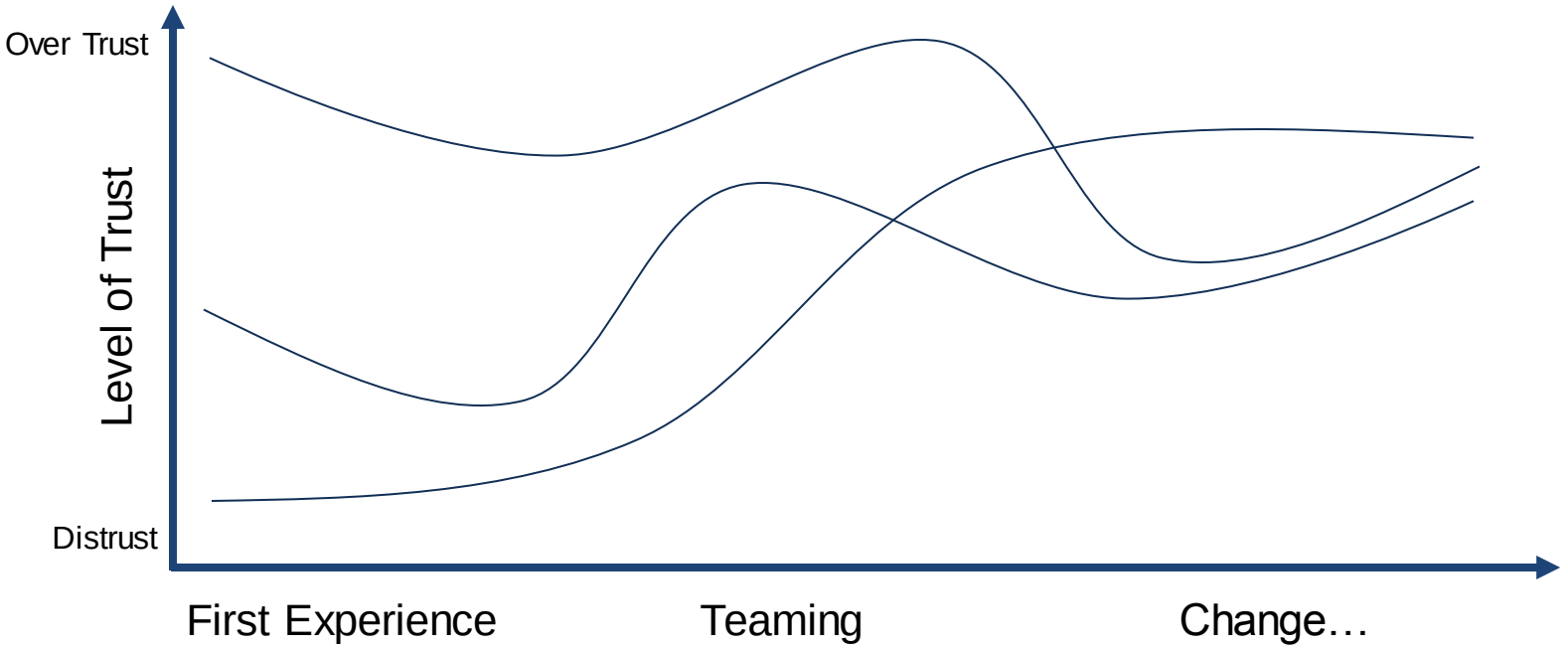
Over Trust
Trust exceeding system capabilities - may lead to misuse.

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

UX research methods and activities to activate curiosity:

- Abusability Testing (Dan Brown)
- "Black Mirror" Episodes (Casey Fiesler)
  (inspired by British dystopian sci-fi tv series of same name)

Speculate about system misuse and abuse

- What are potential unintended/unwanted consequences?

Reward team members for finding ethics bugs (Dr. Ayanna Howard)

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

AINOW INSTITUTE

Microsoft

Google

< >
Montréal Declaration
Responsible AI_
</ >

An initiative of Université de Montréal

U.S. DEPT OF DEFENSE

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-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.

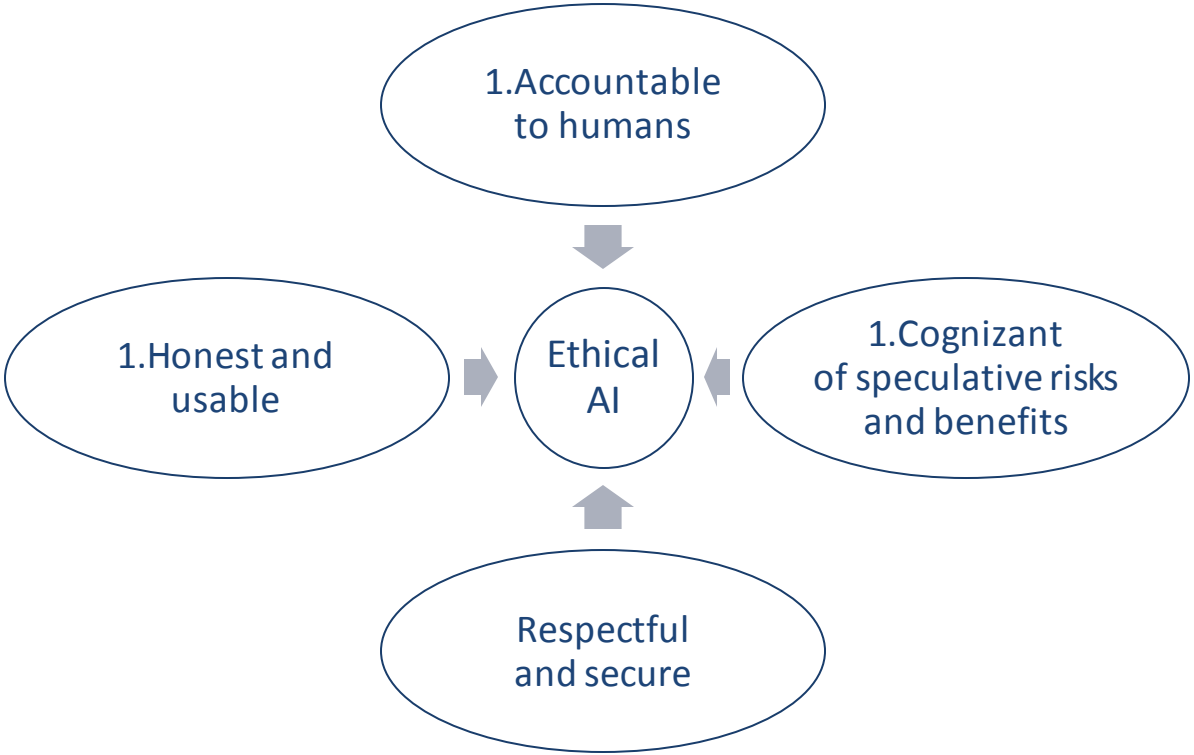
Team Signatures and Date

About the SEI

Contact Us

Checklist and Agreement - Downloadable PDF:
https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620

UX Framework for Designing Trustworthy AI



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

RightStaff Scenario

AI shift scheduling system

Users: Store managers of fast-food restaurants

Goals of RightStaff:

- Faster staffing decisions and scheduling
- Reduced bias of shift selection

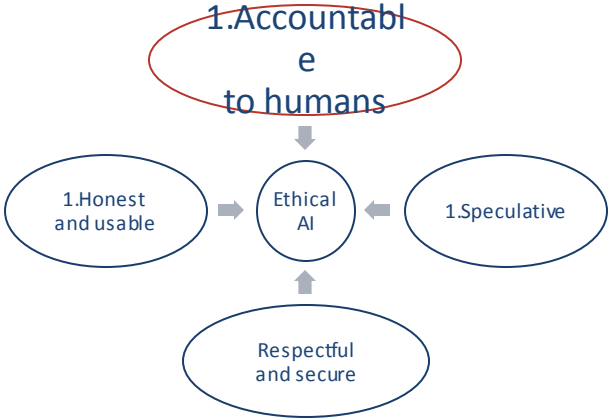
Accountable to Humans

Ensure humans have ultimate control

- Able to monitor and control risk

Human responsibility for final decisions

- Person's life
- Quality of life
- Health
- Reputation



“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 the AI system will be

- explained
- able to be overridden
- appealable and reversible

RightStaff

- Manager able to reschedule people as needed

Responsibilities and limitations explicitly defined

For AI system and human(s)

RightStaff (AI System *or Manager?*)

- Picks employees to schedule?
- Defines shifts?
- Method to integrate new information?
  - Sick time
  - Resignations

Abusability Testing

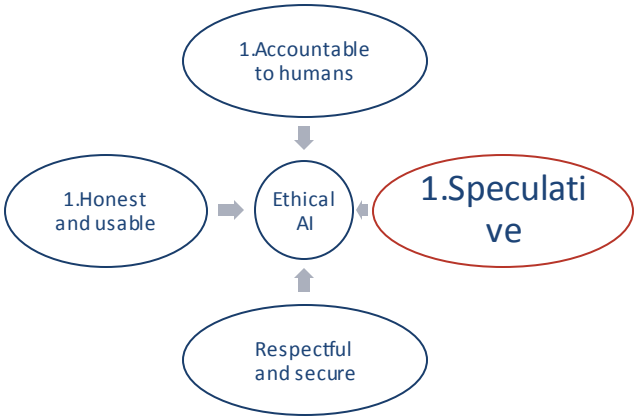
Feature added to enable RightStaff to turn off by itself

- What are limits to functionality?
- How is the situation communicated?
- How could this be abused/misused?
- Implications?
- Risks?

Cognizant of Speculative Risks and Benefits

Identify full range of

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

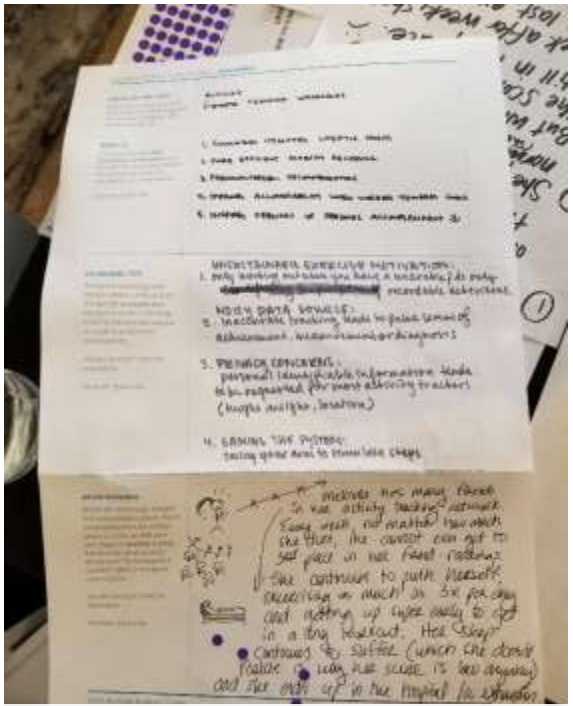


Conduct UX research - activate curiosity

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

“Black Mirror” episode

RightStaff begins prioritizing people with easier schedules

Managers approve these schedules, reinforcing bias

People who were previously discriminated against
are still discriminated against

What else?

Create communication & mitigation plans

Plan for unwanted consequences

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

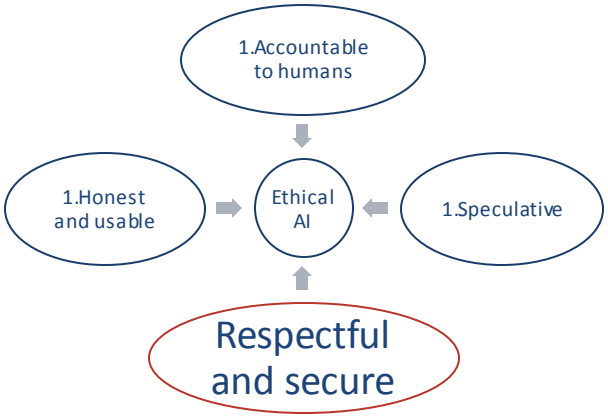
Respectful and Secure

Values of humanity, ethics, equity, fairness, accessibility, diversity and inclusion

Respect privacy and data rights

Make system robust, valid and reliable

Provide understandable security



Respectful and Secure

RightStaff

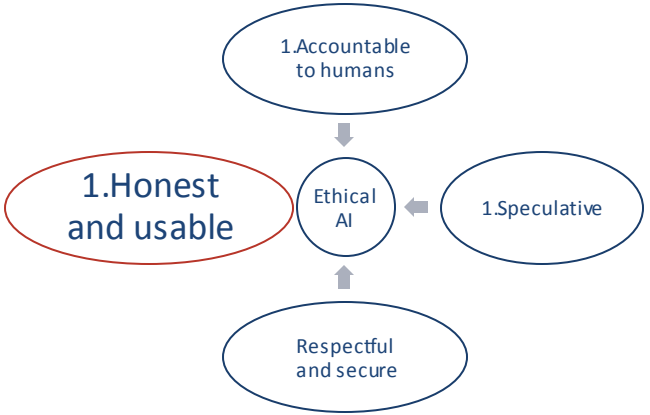
- Who has visibility to reasons for changing schedules?
- How is that information used?
- How is PII* of employees protected?

*PII is Personally Identifiable Information (social security number, address, etc.)

Honest and Usable

Value transparency with the goal of engendering appropriate trust

Explicitly state identity as an AI system



Fair: Unwanted bias in data

Show awareness of known and desirable bias

Acknowledge issues

Overcommunicate on issues

RightStaff

- System built to reduce the known bias in existing data
- Make it easy to report bias (or prevent it)

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

Challenges

Need more speculative activities

Engage people in this hard and necessary work

A challenge - building on our experiences with UX/HCI and accessibility…

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

Methods, Mechanisms, and Mindsets

Engage in Critical Oversight

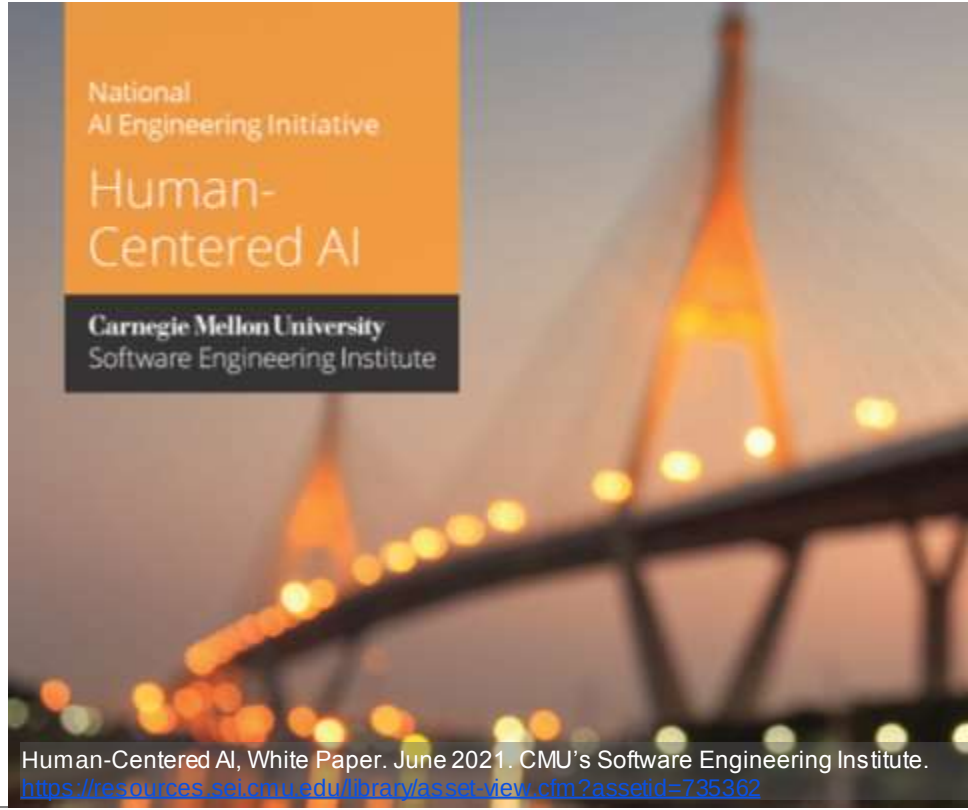
Engage in critical oversight

"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

Data transparency

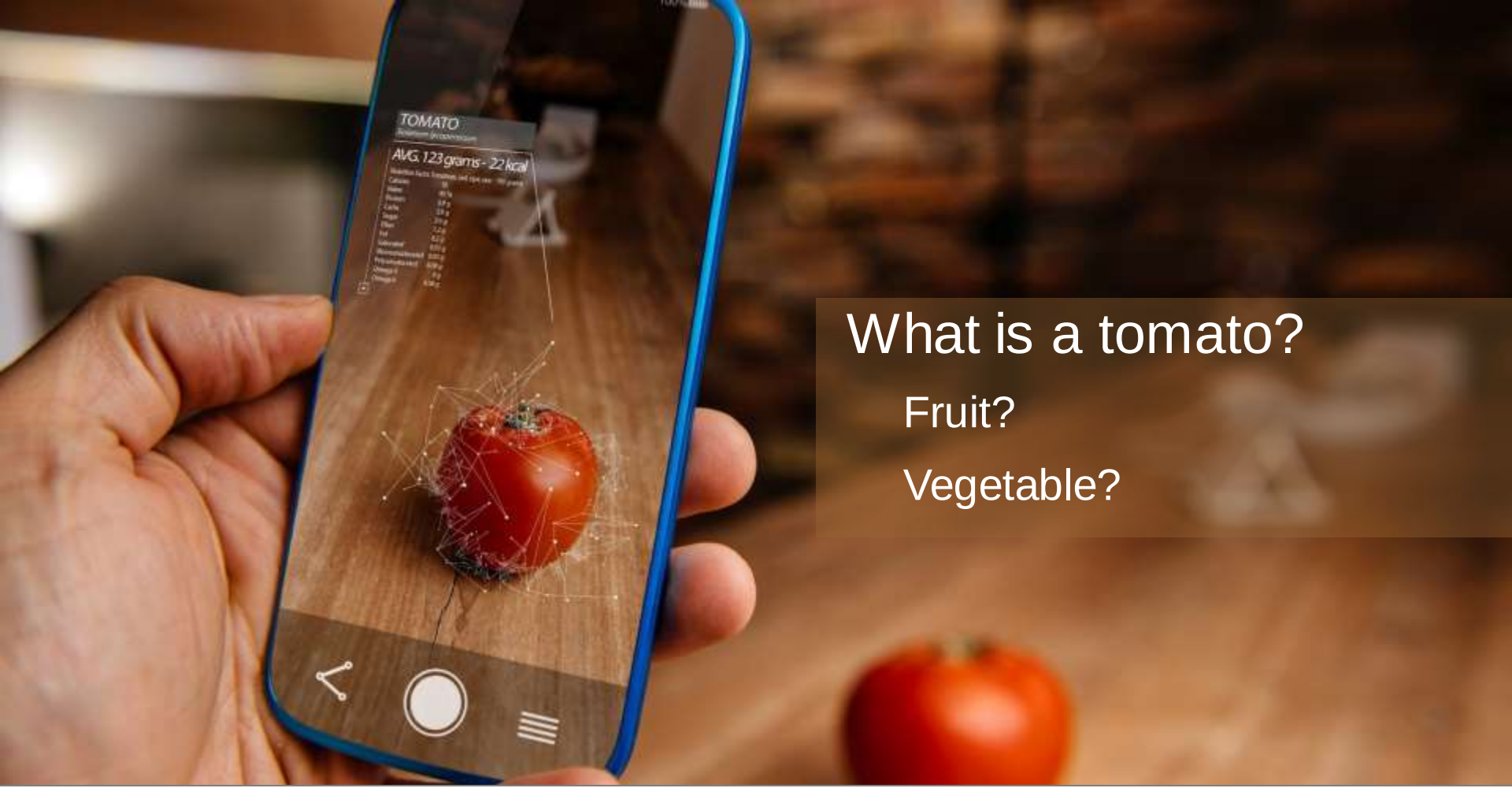
Team must understand the data

- What is it about?
- Provenance, creator's motivation

Resources

- Datasheets for Datasets[1]
- Model Cards for ML systems[2]

1. T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. Datasheets for Datasets. The latest version of this paper can be found online at https://arxiv.org/abs/1803.09010
2. M. Mitchell et al., "Model Cards for Model Reporting," Proc. Conf. Fairness Account. Transpar., pp. 220–229, Jan. 2019, doi: 10.1145/3287560.3287596



What is a tomato?

Fruit?

Vegetable?

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

Bias in data, algorithm selection, and training

Unintended and purposeful bias

Misuse and abuse of the system

Understand inherent bias and amount of variance – data:

- Motivation
- Composition
- Collection process
- Recommended uses, etc.

Goal: Transparency and accountability.

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

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

Regular Auditing

Dynamic systems (not stable)

Continuous human oversight required

- Probe with hypothetical cases
- Checks for bias, brittleness or potential distribution shift
- Access history of system operation and usage*

*Consider ethical principles for data collection.

Leaders must establish psychological safety

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

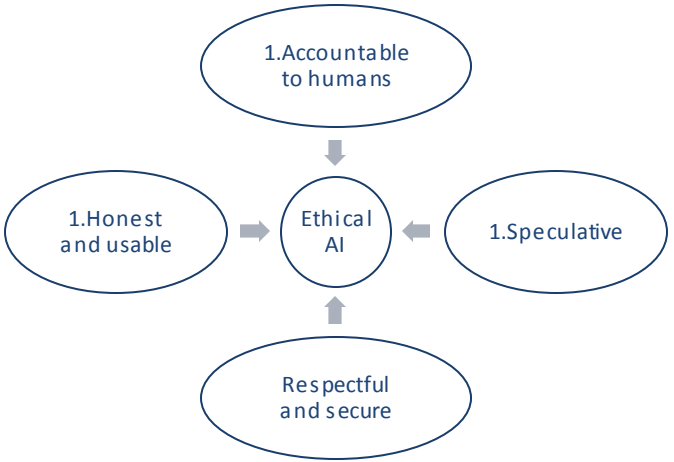
We aren’t perfect, AI won’t be perfect

“AI will ensure appropriate human judgement and not replace it”
- Defense Innovation Board. 2019

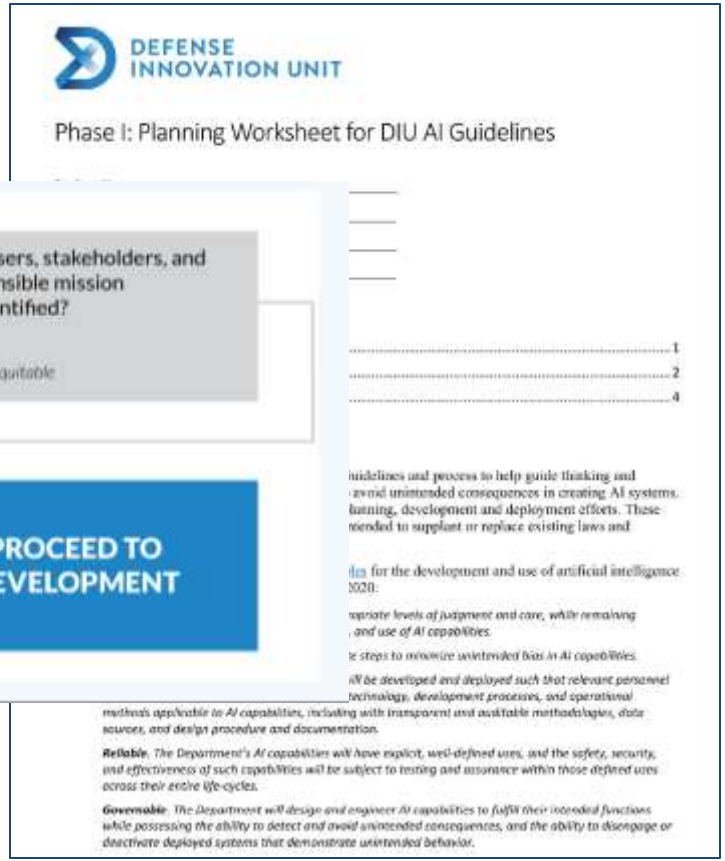
Empower diverse teams, inclusive environments

Encourage deep conversations

Activate curiosity; be speculative; imaginative



RAI Worksheets, Report, and Workshops


DEFENSE INNOVATION UNIT

Phase I: Planning Worksheet for DIU AI Guidelines

https://www.diu.mil/responsible-ai-guidelines

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