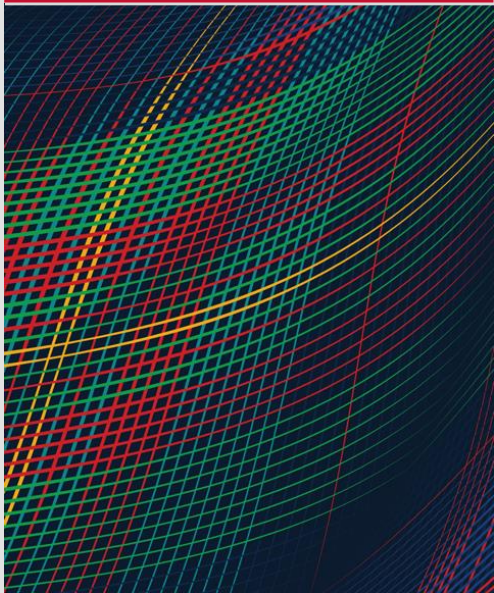
Carnegie Mellon University
Software Engineering Institute

Operationalizing Responsible AI

NOVEMBER 8, 2023 - AI/ML TEM

Carol J. Smith, Principal Research Scientist and Trust Lab Lead
SEI AI Division



Copyright Statement

DM23-2162

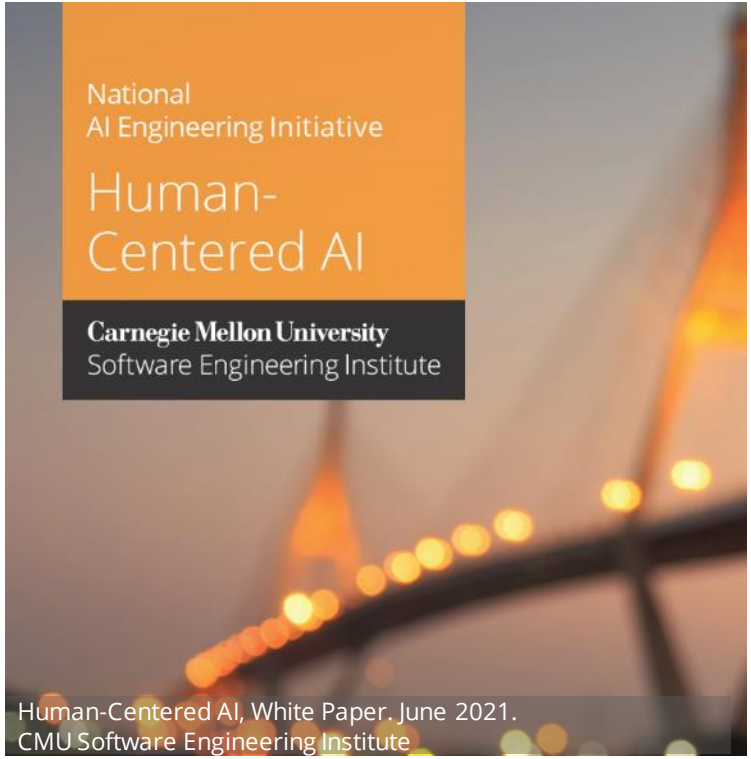
Design to work with, and for, people

Carnegie Mellon University Software Engineering Institute

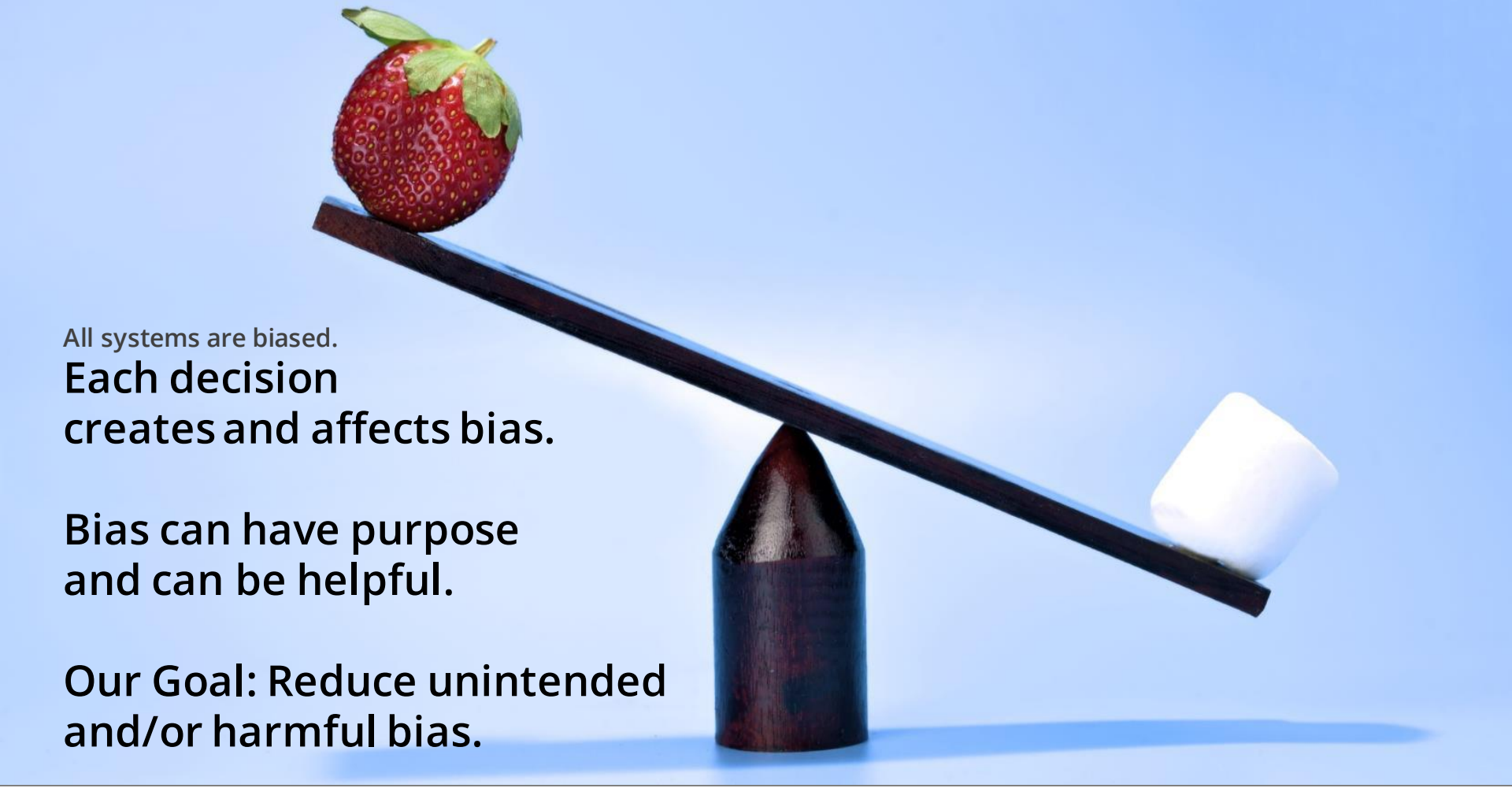
Provide trustworthy interactions.

We must design AI systems to:

- be accountable to humans
- identify and explain risks
- be respectful, honest, and usable



Human-Centered AI, White Paper. June 2021.
CMU Software Engineering Institute



All systems are biased.

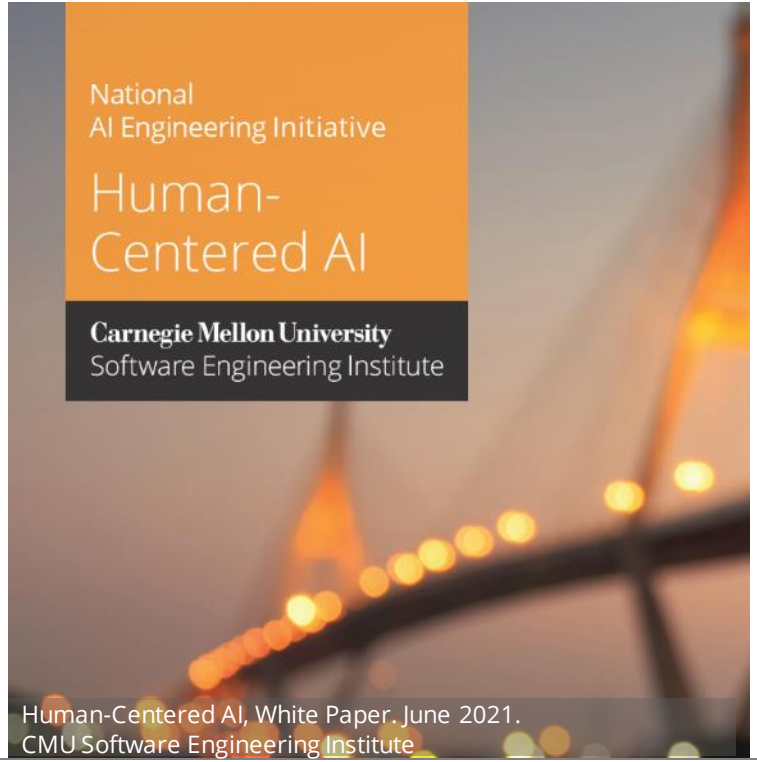
Each decision creates and affects bias.

Bias can have purpose and can be helpful.

Our Goal: Reduce unintended and/or harmful bias.

Trustworthy, Human-Centered, and Responsible AI

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



Human-Centered AI, White Paper. June 2021.
CMU Software Engineering Institute

Sense changes over time

Understand Complexity of Context

Carnegie Mellon University
Software Engineering Institute

Identify Sources of Bias and Complexity

Environment
People
Information
System capabilities in context

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

Observe Use of System in Environment

Carnegie Mellon University Software Engineering Institute

Who will use the system?
What are their needs?
What do they need to do?

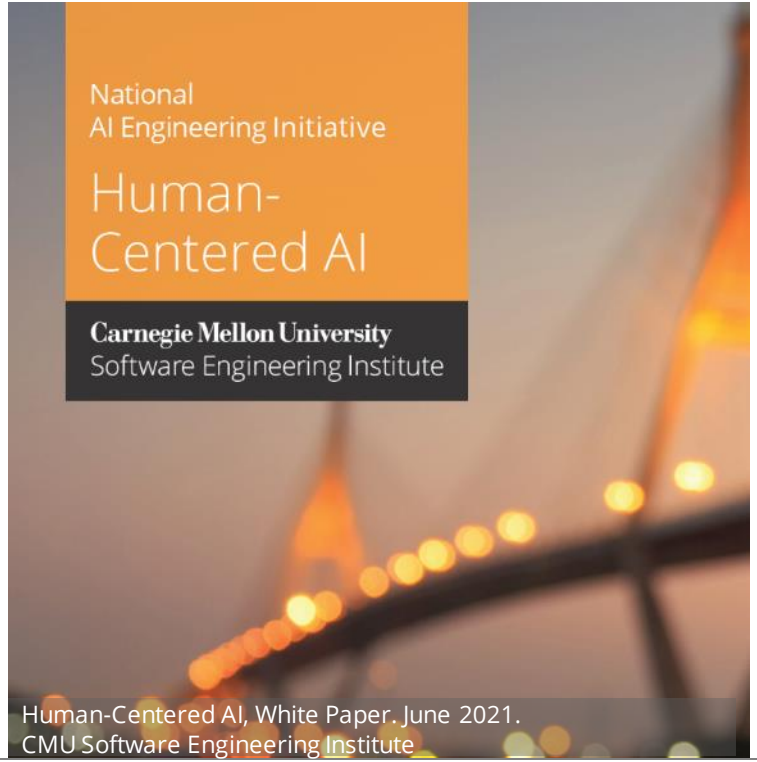
Conduct Research
Observations,
Interviews


U.S. Air Force photo by Master Sgt. Barry Loo

Exchanges and Context

How do human and AI:

- share information, capabilities, situational awareness?
- learn when shifts have occurred?
- manage change over time?
- adapt and evolve based on dynamic contexts?



Human-Centered AI, White Paper. June 2021.
CMU Software Engineering Institute

Speculation Keeps People Safe

U.S. Air Force photo by Airman 1st Class Ethan Sherwood. Goodfellow Air Force Base, TX, United States. https://www.dvidshub.net/image/6443325/drones-goodfellow

Activate Curiosity

Carnegie Mellon University Software Engineering Institute

Speculate about unintended and/or unwanted consequences.

What are potential

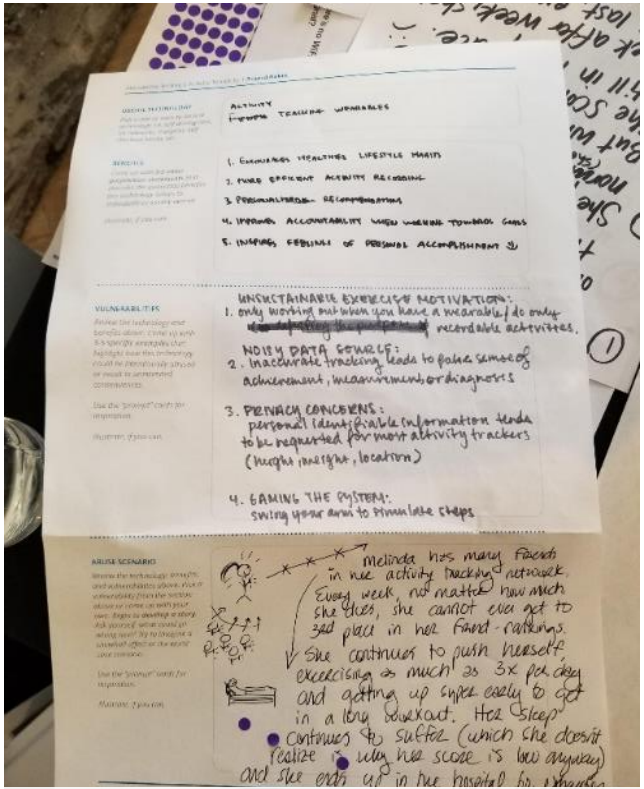
- severe misuse and abuse?
- negative consequences for people in marginalized groups?

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

Abusability Testing

1) Value proposition
2) Vulnerabilities
3) Abuse scenario

Provocation via prompts



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 , Pittsburgh IxDA Chapter for World Interaction Design Day 2019.

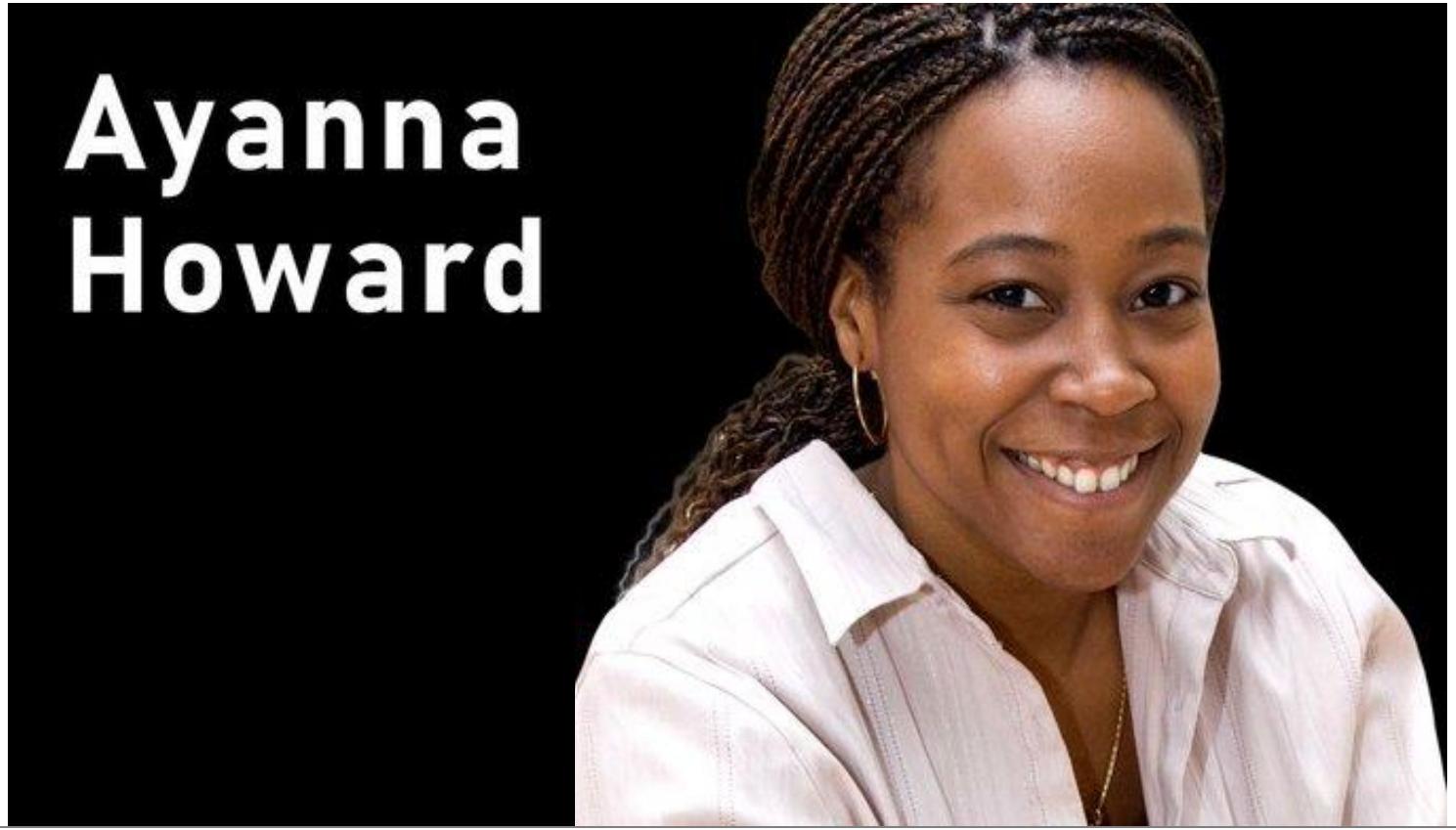
3Q-Do No Harm Framework

Carnegie Mellon University Software Engineering Institute

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

Reward team members for finding ethics bugs



Card Game: What Could Go Wrong?

Carnegie Mellon University Software Engineering Institute

Foster conversations around potential challenges and issues with complex technologies.

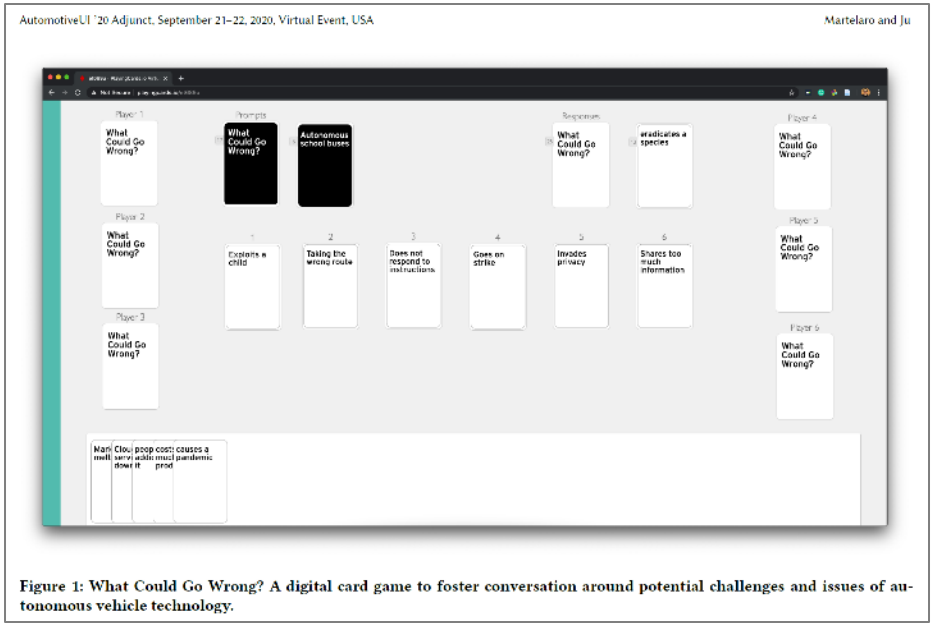


Figure 1: What Could Go Wrong? A digital card game to foster conversation around potential challenges and issues of autonomous vehicle technology.

Nikolas Martelaro and Wendy Ju. 2020. What Could Go Wrong? Exploring the Downsides of Autonomous Vehicles. In 12th Internatio nal 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

Understand context

Carnegie Mellon University Software Engineering Institute

Human-centered research to identify
- Complexity and sources
- Changes over time

Inform and support designs that provide evidence to users.

Challenges

- Demystifying AI
- Creating more speculative activities
- Engaging teams in this hard and necessary work

Development of tools, processes, and practices

Human-Machine Teaming

Carnegie Mellon University
Software Engineering Institute

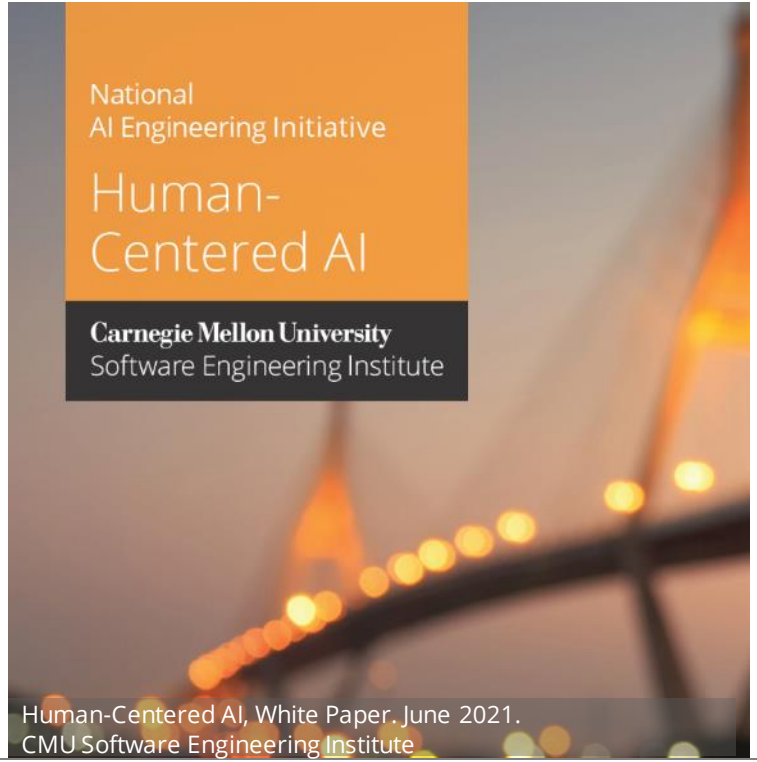
Teaming requires trustworthy AI

Carnegie Mellon University Software Engineering Institute

Capabilities (and limitations) are explained – transparency.

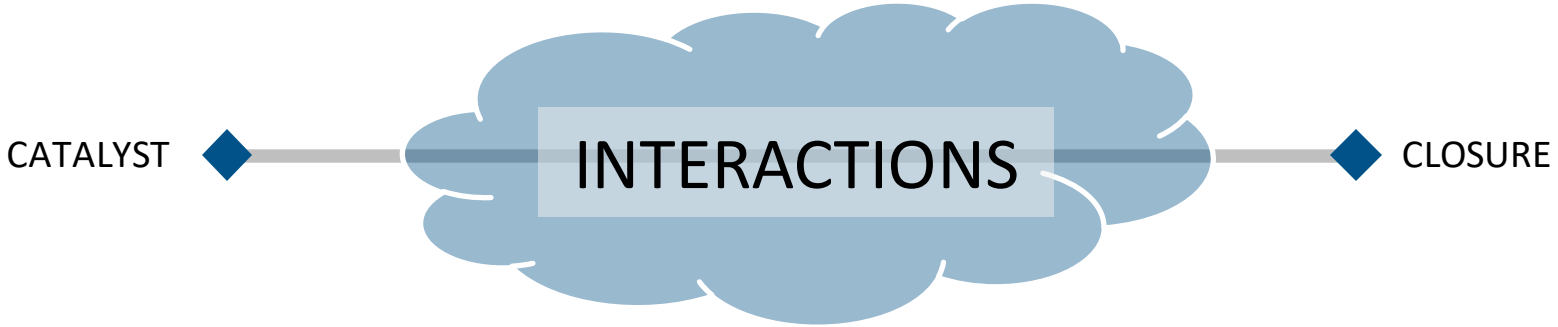
Continuous monitoring and oversight are prioritized.

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



Human-Centered AI, White Paper. June 2021.
CMU Software Engineering Institute

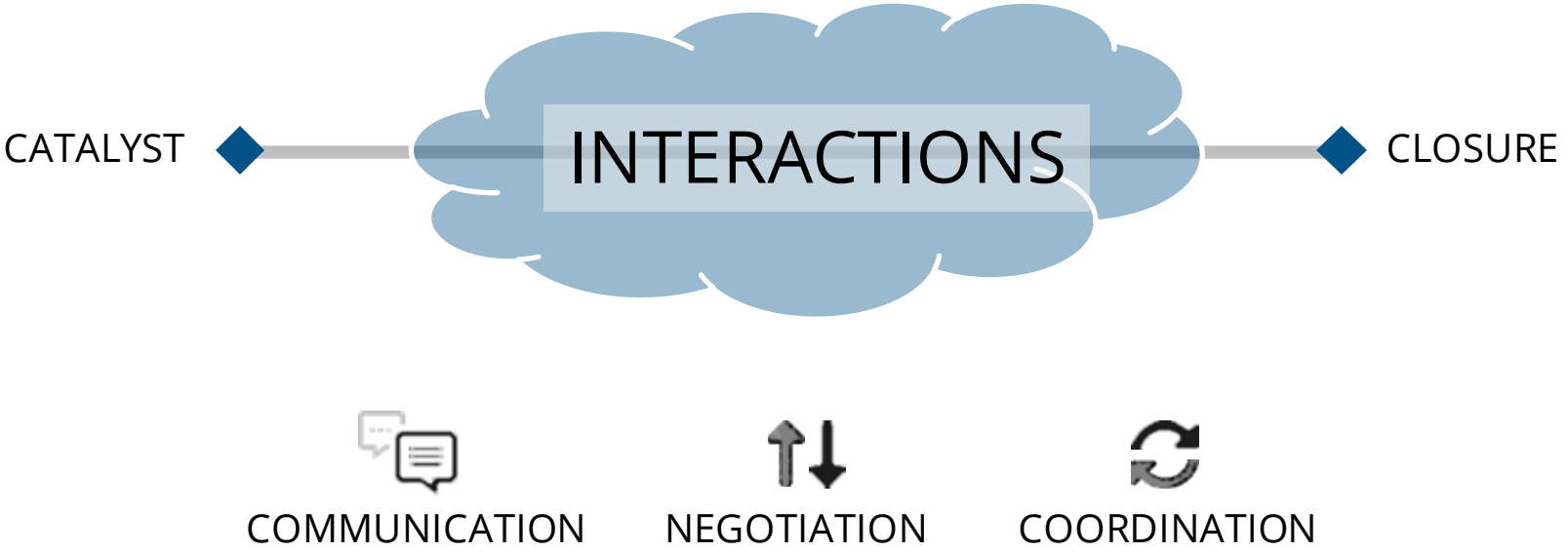
Identify collaborative activities and interactions



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

Medical treatments - decision support

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)

Recognize human strengths

Carnegie Mellon University Software Engineering Institute

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

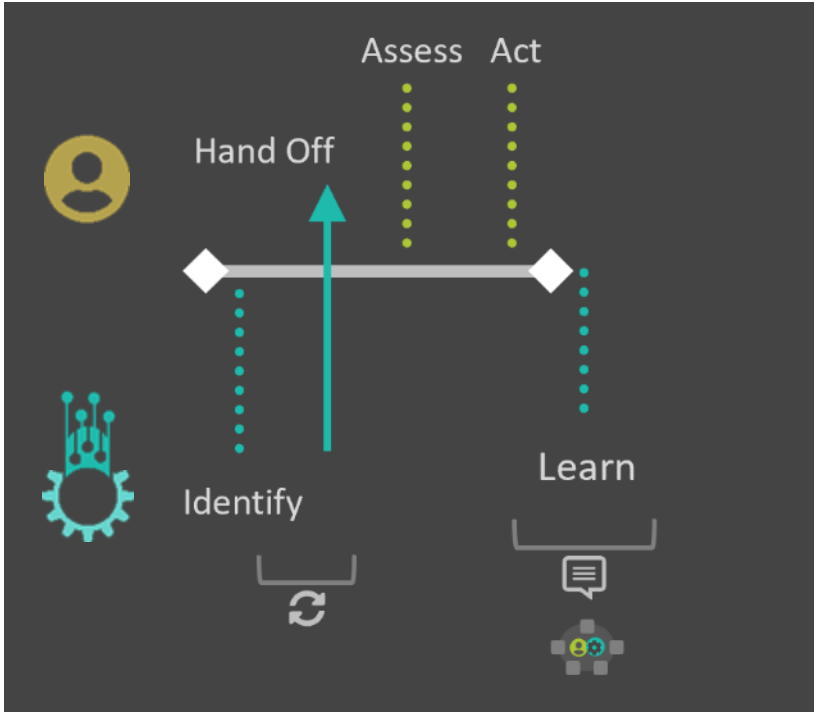
Amanda Muller and Carol Smith. 2022. Perceptions of Function Allocation between Humans and AI-Enabled Systems. UXPA 2022 (pre-print). https://uxpa2022.org/sessions/perceptions-of-function-allocation-between-humans-and-ai-enabled-systems/

Consider Responsibility

Responsibilities explicitly defined between people and systems.

Significant decisions made by system

- explained
- able to be overridden
- appealable and reversible



How IAs Can Shape the Future of Human-AI Collaboration. Carol Smith and Duane Degler. Presented on April 28-30, 2021 at IAC21.

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

Adopt Technology Ethics

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

IMMEDIATE RELEASE

DOD Adopts Ethical Principles for Artificial Intelligence

FEB. 24, 2020

The U.S. Department of Defense officially adopted a series of ethical principles for the use of Artificial Intelligence today following recommendations provided to Secretary of Defense Dr. Mark T. Esper by the Defense Innovation Board last October.

The recommendations came after 15 months of consultation with leading AI experts in commercial industry, government, academia and the American public that resulted in a rigorous process of feedback and analysis among the nation's leading AI experts with multiple venues for public input and comment. The adoption of AI ethical principles aligns with the DOD AI strategy objective directing the U.S. military lead in AI ethics and the lawful use of AI systems.

"The United States, together with our allies and partners, must accelerate the adoption of AI and lead in its national security applications to maintain our strategic position, prevail on future battlefields, and safeguard the rules-based international order," said Secretary Esper. "AI technology will change much about the battlefield of the future, but nothing will change America's steadfast commitment to responsible and lawful behavior. The adoption of AI ethical principles will enhance the department's commitment to upholding the highest ethical standards as outlined in the DOD AI Strategy, while embracing the U.S. military's strong history of applying rigorous testing and fielding standards for technology innovations."

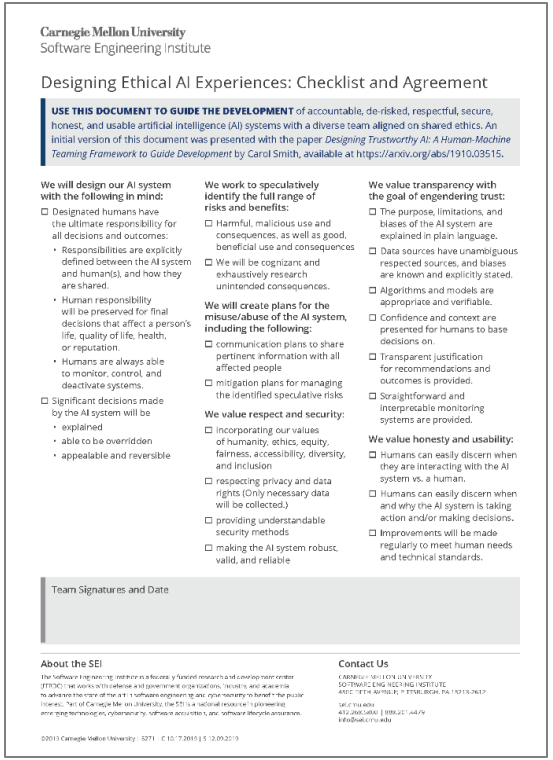
DoD Memorandum, "Artificial Intelligence Ethical Principles for the Department of Defense." (Feb 2020)

Use Frameworks to Guide Responsible AI

Carnegie Mellon University Software Engineering Institute

Pair checklists with technical ethics as prompts for conversations

- 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 final decisions that affect a person's life, quality of life, health, or reputation.
  - Humans are always able to monitor, control, and deactivate systems.
- ☐ Significant decisions made by the AI system will be
  - explained
  - able to be overridden
  - appealable and reversible

We work to speculatively identify the full range of risks and benefits:

- ☐ Harmful, malicious use and consequences, as well as good, beneficial use and consequences
- ☐ We will be cognizant and exhaustively research unintended consequences.

We will create plans for the misuse/abuse of the AI system, including the following:

- ☐ communication plans to share pertinent information with 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 context are presented 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

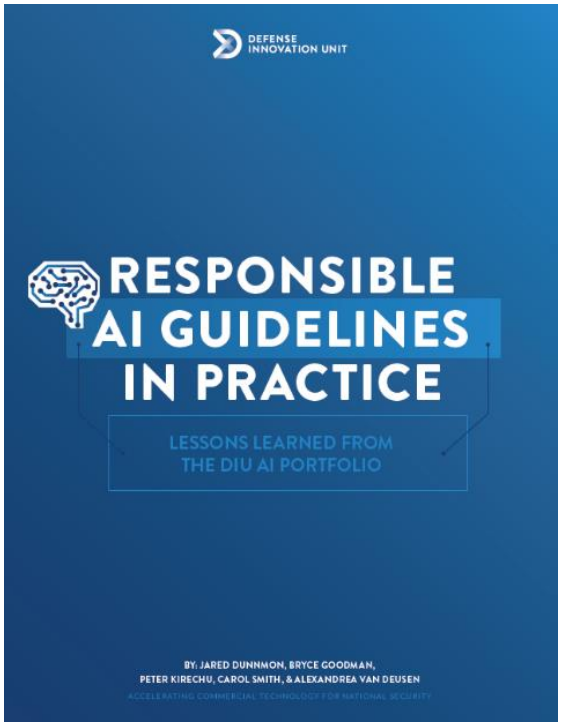
About the SEI

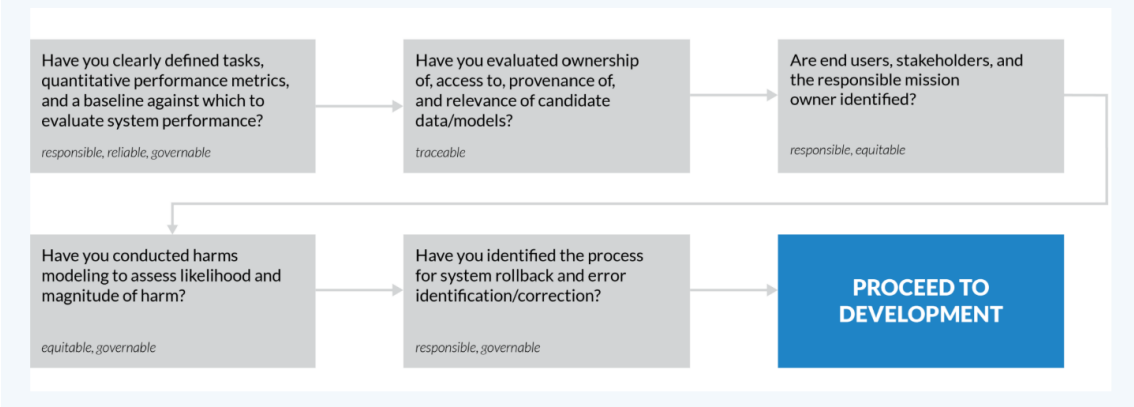
Contact Us

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

Defense Innovation Unit

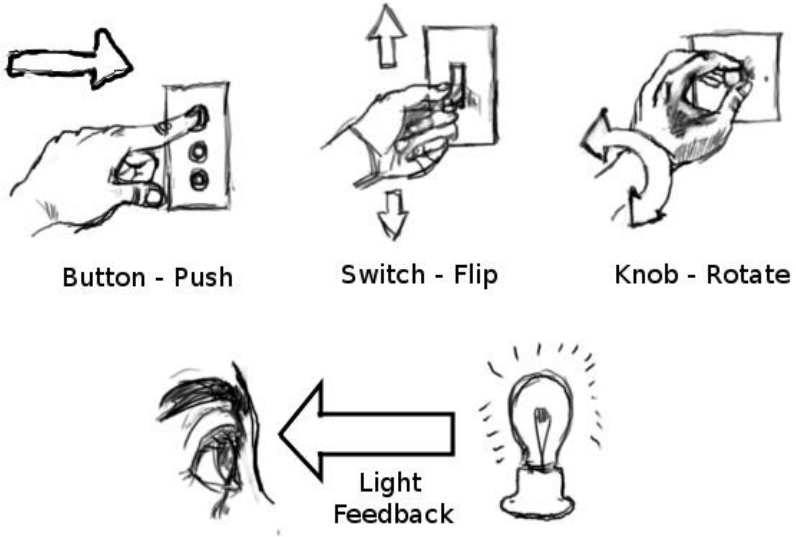
RAI Report, Guidelines, Worksheets, and Workshops

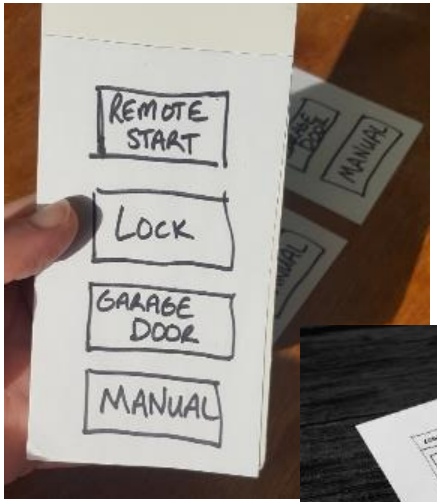


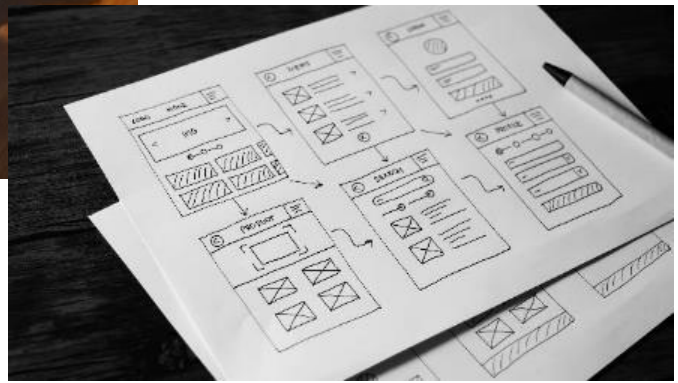


Defense Innovation Unit. Artificial Intelligence Portfolio, Responsible AI Guidelines. https://www.diu.mil/responsible-ai-guidelines

Make Informed Decisions + Prototypes







Drawings of Affordance: http://paaralan.blogspot.com/2010/09/affordance-and-educational-games.html

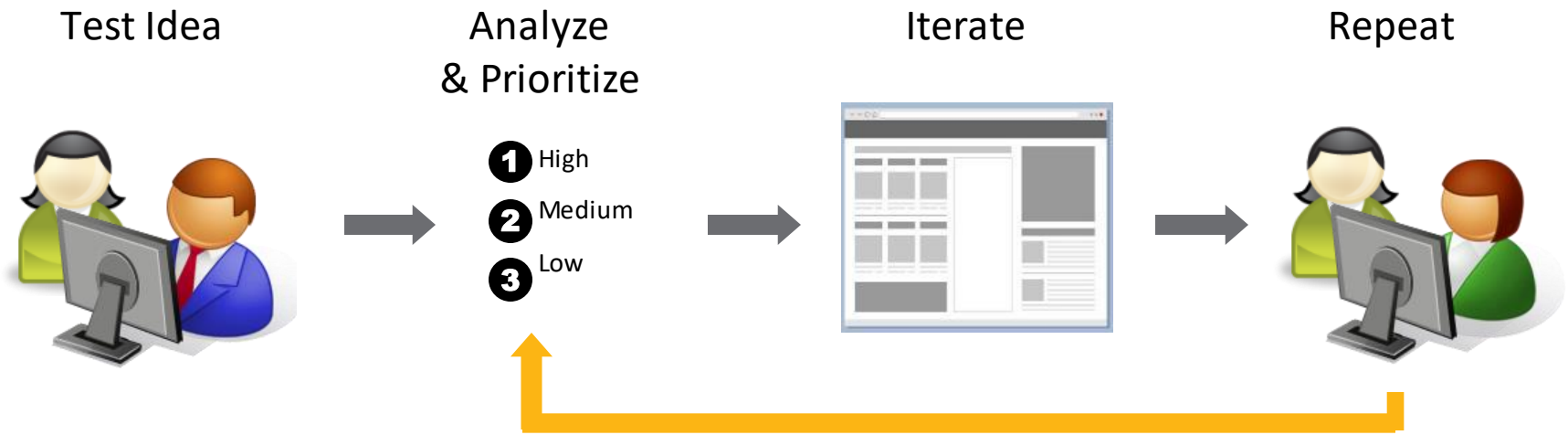
Test System with Human Teammates

Prototype prior to significant development.

Are the human-machine teams able to complete tasks?

Evidence of…
- Improved efficiency/productivity?
- fewer errors?
- reduced training time?
- improved acceptance?

Iterative Cycles: Feedback and Improvement



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

Challenges

- Considerations for change over course of use
- Measurements of trustworthiness
- AI Systems are not fully able to team with humans yet,
  - but we need to be ready!

Methods, Mechanisms, and Mindsets

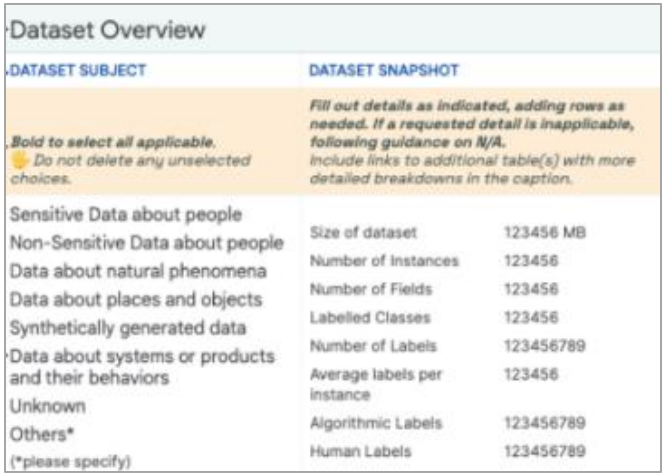
Engage in Critical Oversight

Carnegie Mellon University
Software Engineering Institute

Mitigate Unwanted Bias Holistically

Understand the complete picture of data-derived bias

- Data creator's funding, motivation and collection process
- Inherent bias and amount of variance in collected data
- Rationale for data inclusion, and what was excluded
- Recommended uses of dataset
- Sensitivity of data



Dataset Overview

DATASET SUBJECT	DATASET SNAPSHOT	
Bold to select all applicable. *Do not delete any unselected choices.*	*Fill out details as indicated, adding rows as needed. If a requested detail is inapplicable, following guidance on N/A.* *Include links to additional table(s) with more detailed breakdowns in the caption.*	
Sensitive Data about people	Size of dataset	123456 MB
Non-Sensitive Data about people	Number of Instances	123456
Data about natural phenomena	Number of Fields	123456
Data about places and objects	Labelled Classes	123456
Synthetically generated data	Number of Labels	123456789
Data about systems or products and their behaviors	Average labels per instance	123456
Unknown	Algorithmic Labels	123456789
Others*	Human Labels	123456789
(*please specify)		

Sample data card, source: Pushkarna et al., 2022

Humans Must Constantly Monitor AI Systems (for now)

AI systems are

- Not “stable” like typical software
- An expensive investment

- Proactively consider risks
- Probe with hypothetical cases
- Checks for bias, brittleness, potential distribution shift

Implementation and Continuous Oversight

Plan for long term…

AI are great at finding patterns, but

- can learn the wrong pattern
- connect the wrong data, and are
- dynamic (not like previous software).

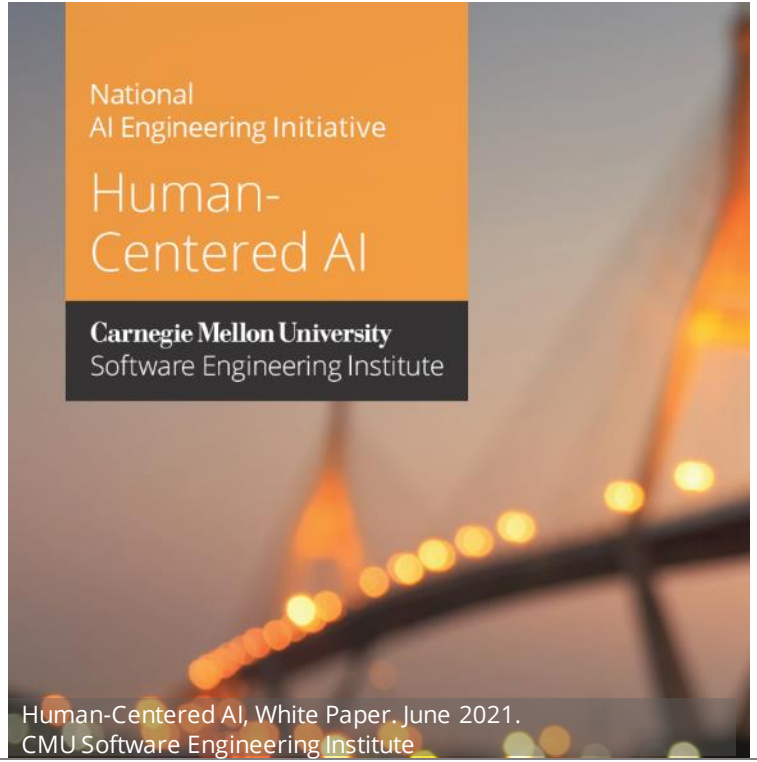
Nacho Kamenov & Humans in the Loop / Better Images of AI /
A trainer instructing a data annotator on how to label images / CC-BY 4.0

Challenges

- Better tools to evaluate data.
- Define standard methods and processes for evaluating system outcomes.
- Methods for continuous oversight.

Trustworthy, Human-Centered, and Responsible AI

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



Human-Centered AI, White Paper. June 2021.
CMU Software Engineering Institute

Carnegie Mellon University
Software Engineering Institute

Responsible AI systems are designed to work with, and for, people

Carol J. Smith, AI Division
cjsmith@sei.cmu.edu