Implementing Responsible, Ethical, and Human-Centered AI

Alex Van Deusen and Carol J. Smith
AI Division, CMU SEI
arvandeusen@sei.cmu.edu and cjsmith@sei.cmu.edu

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.

How can teams harness the **power** of AI systems and design them to be **valuable** to humans?

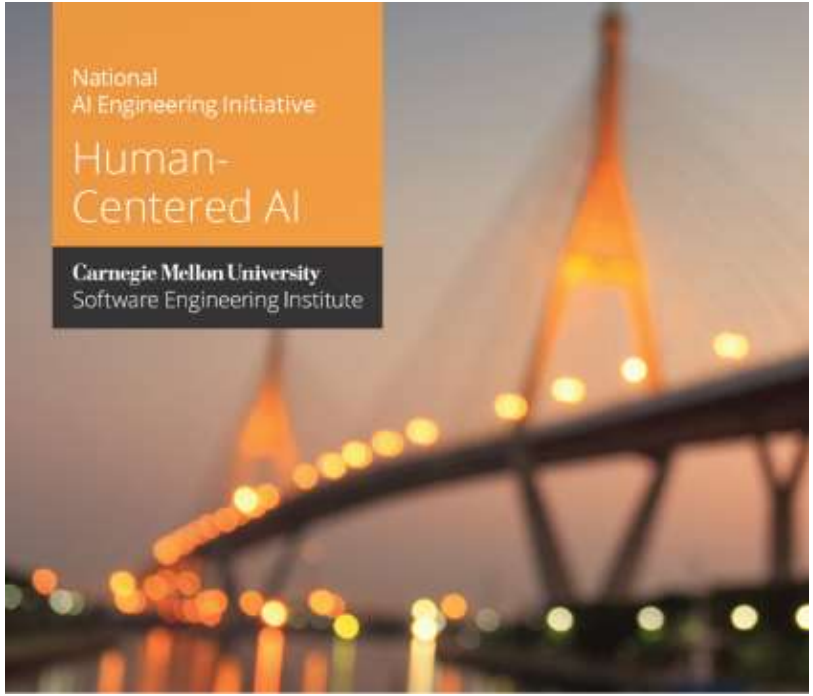
© 2022 Carnegie Mellon University

[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.

Advancing human-centered AI

Effective implementations minimize unintended consequences

- Understand complexity of context
- Design for human-machine teaming
- Engage in critical oversight

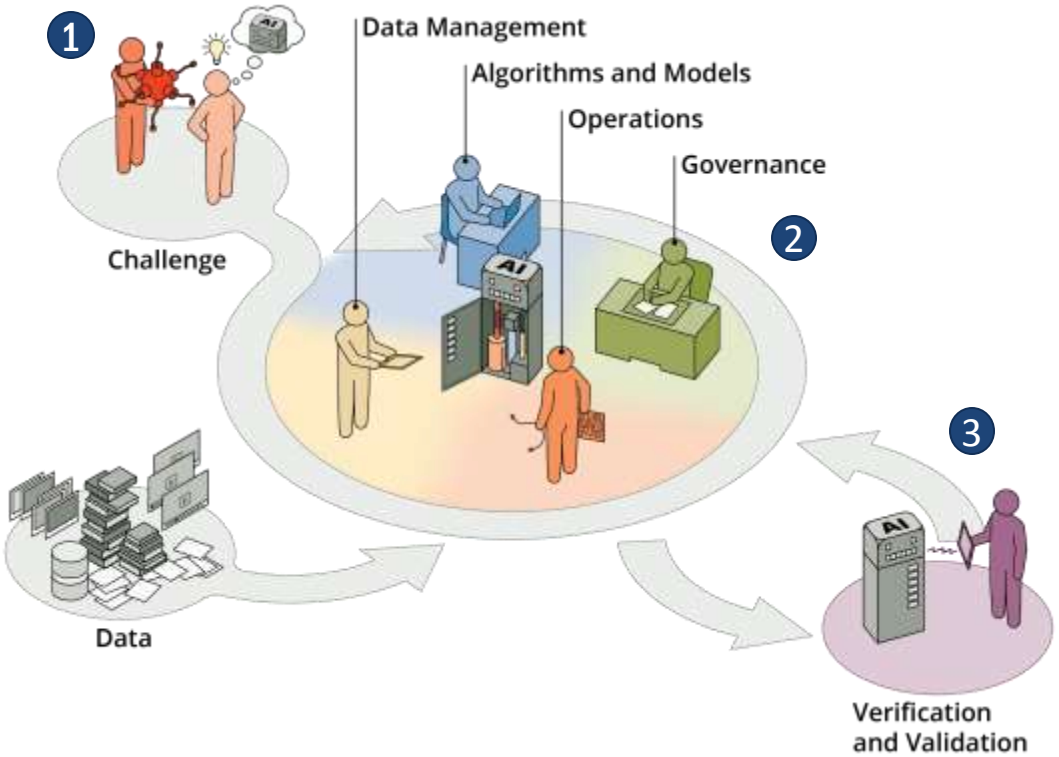


Human-Centered AI, Software Engineering Institute: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362

© 2022 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 Systems and Human-Machine Teaming



1. **Understand complexity of context:** focus on user needs
2. **Design for human-machine teaming:** align on technical ethics
3. **Engage in critical oversight:** Make AI interpretable, understandable, verifiable

© 2022 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.

Sensing changes over time

Understand Complexity of Context

© 2022 Carnegie Mellon University

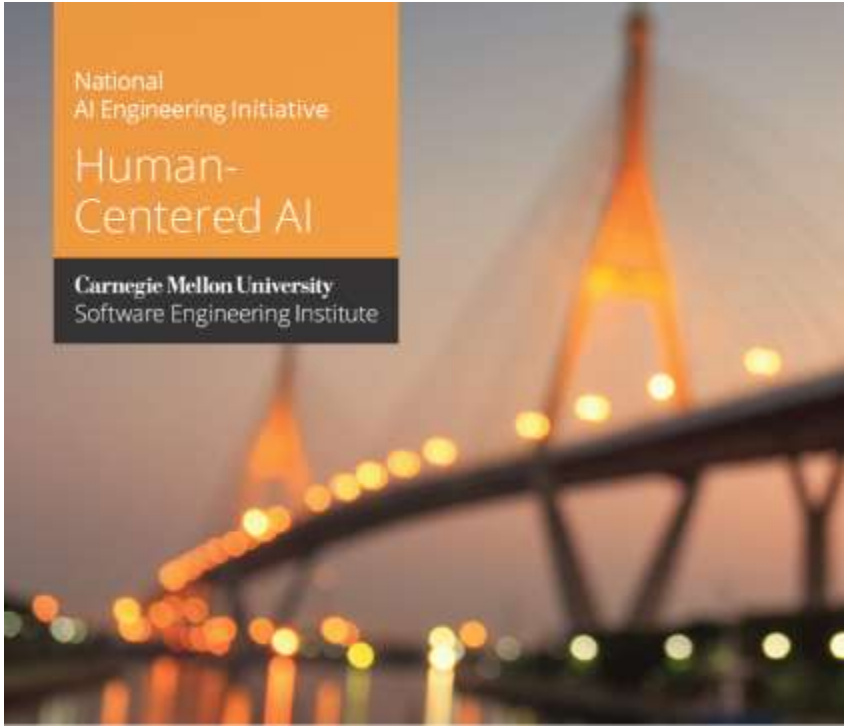
[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

Understanding context and sensing changes over time

- What is the desired system outcome?
- Human and contextual factors affect outcome

Does the human-machine team:

- Learn when shifts in context have occurred?
- Maintain clarity around operational intent?
- Adapt and evolve based on dynamic contexts and user needs?



Human-Centered AI, Software Engineering Institute: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362

© 2022 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.

Complexity of context

- Environmental
- Human
- AI system capabilities
- Information

© 2022 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.

Define the desired outcome

- What problem are you solving?
- For whom?
- What will help them?
- What kind of improvements are expected?
- What might a machine do better or faster?
- What happens if the AI system is *not* used?
- What is *not* going to be improved (out of scope)?

© 2022 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.

Example: Semi-autonomous vehicles

Adaptability to change

- Road conditions
- Weather
- Desired route
- Street painting
- Emergent situations

© 2022 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.

Understanding context of use

- Conduct human-centered research with primary users prior to development efforts
  - Identify humans' actual objectives
  - Sense how objectives evolve, and continuously improve approximations of desired outcomes
- Integrate insights from human-centered research and translate into
  - Design and operations of system
  - AI System Reward functions
- Research needed on opportunities and mechanisms to collect information and sense or infer user intention from other parts of larger system

© 2022 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.

Developing tools, processes, and practices

Designing for Human-Machine Teaming

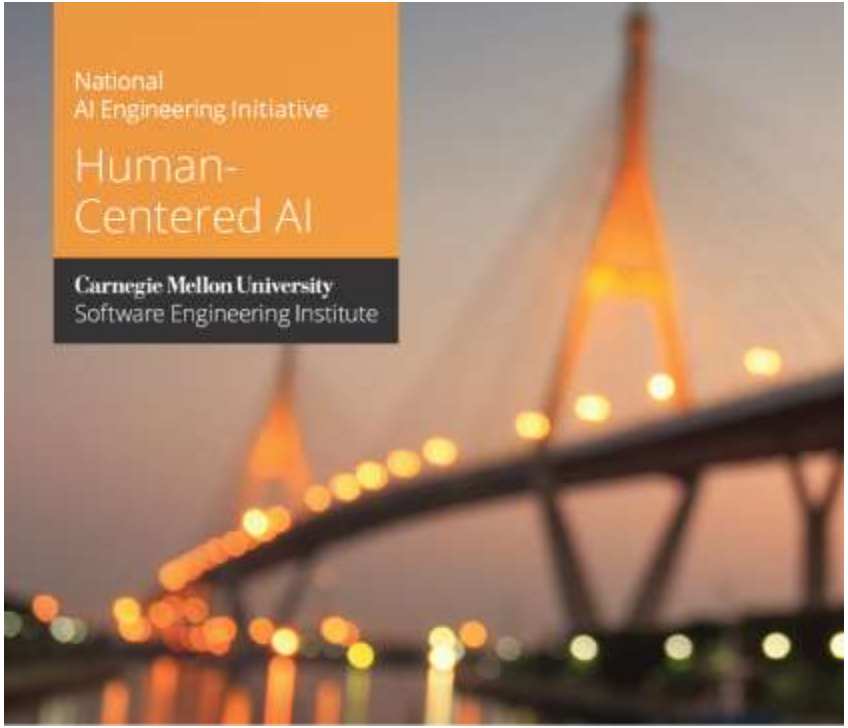
© 2022 Carnegie Mellon University

[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.

Facilitating interdependence

Human and machine complimenting each other

- Requires primary user to interact with and understand system
- Gaining *calibrated* level of trust
- AI system designed provide transparency regarding its limitations



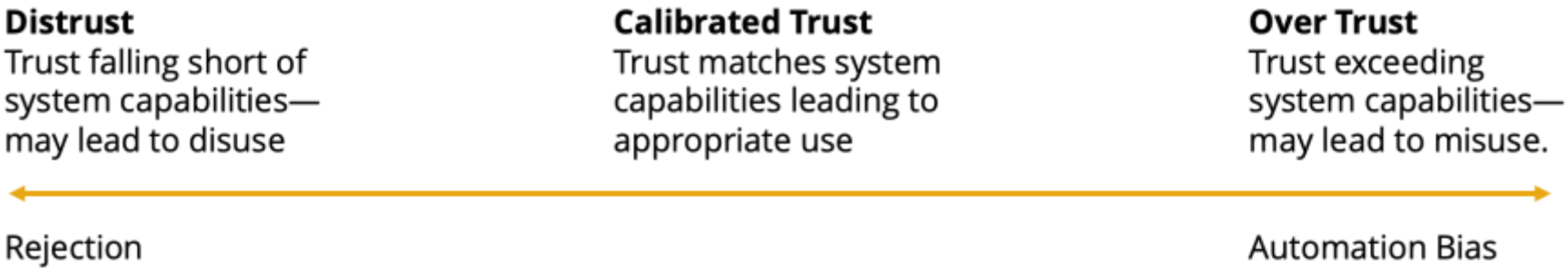
Human-Centered AI, Software Engineering Institute: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362

© 2022 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 available evidence of system's capability and integrity.

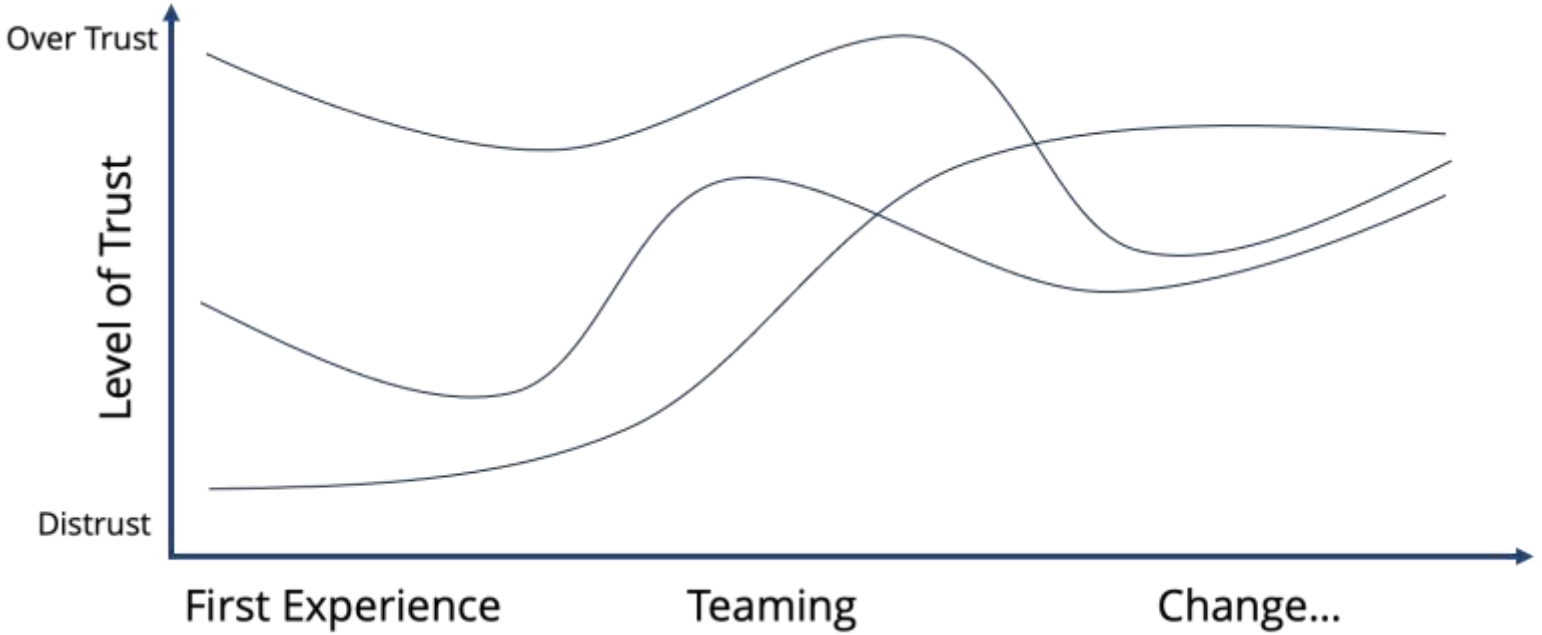


Bobbie Seppelt and John Lee. 2012. Human Factors and Ergonomics in Automation Design. In Handbook of Human Factors and Ergonomics (Fourth Edition) Chapter 59. Wiley.
DOI: https://doi.org/10.1002/9781118131350.ch59

© 2022 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

© 2022 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

Incorporate user experience (UX) research and human-computer interaction (HCI) methods to activate curiosity.

- Speculate about misuse and abuse.
- Identify potential unintended/unwanted consequences.
- Potential severe abuse and consequences.
- Perspectives of people in frequently marginalized groups.

© 2022 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.

Share understanding

Converse on 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

© 2022 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.

Integrate technical ethics

- Pick a set of ethics that works well for you and your team, adjust as necessary
- Pair with checklists and other RAI tools
  - 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 this 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 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

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

© 2022 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.

Designing for Human-Machine Teaming

- Design AI systems to provide transparency regarding AI limitations - boundaries and unfamiliar scenarios.
- Humans will gain *calibrated* levels of trust.
- Speculate about misuse and abuses
- Prevent or plan to mitigate situation.

© 2022 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

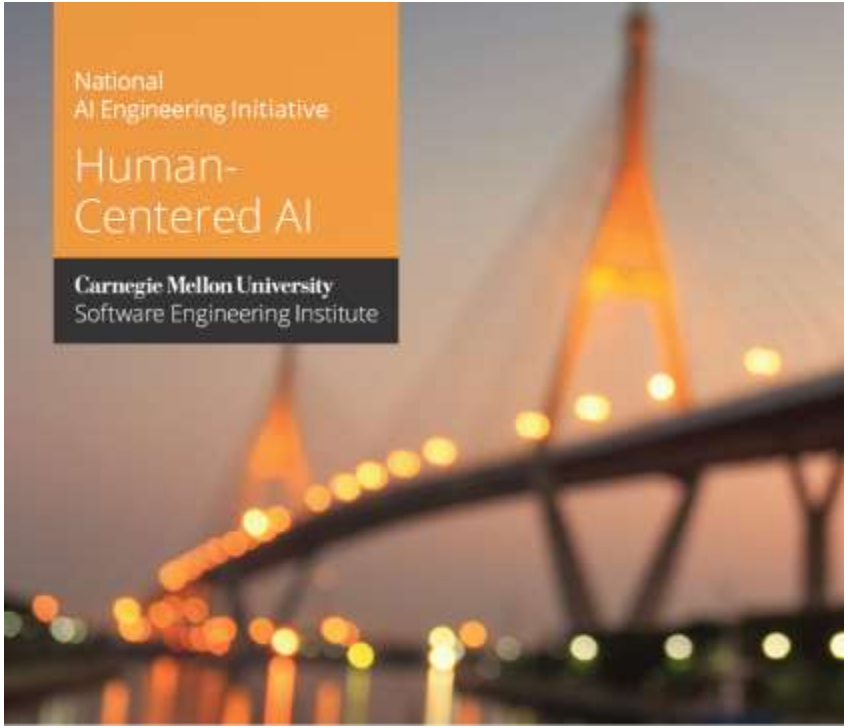
© 2022 Carnegie Mellon University

[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

AI systems learn through data and observations

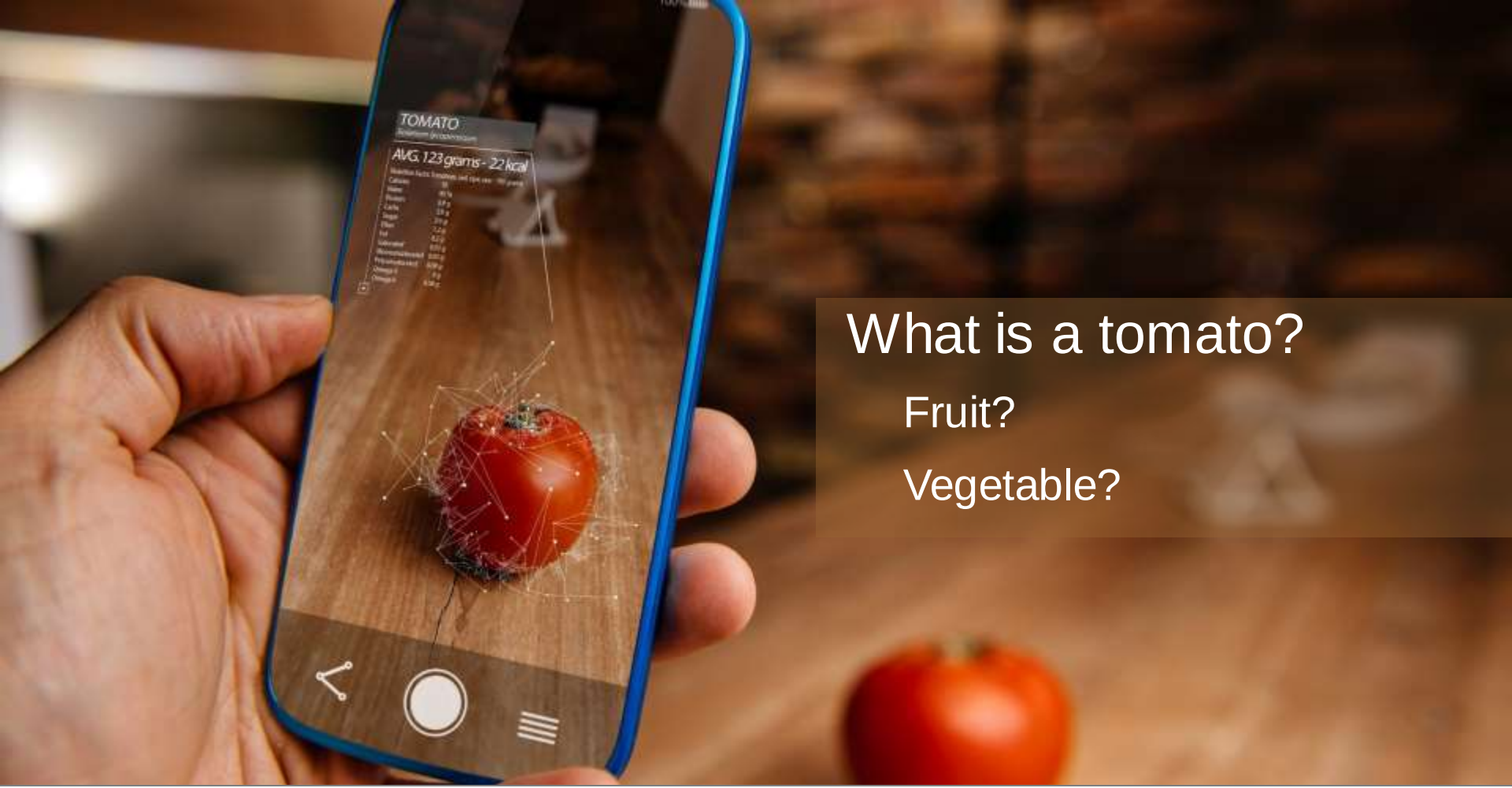
- Continuous human oversight
- Proactively identify risks of bias, misuse, abuse, and unintended consequences



Human-Centered AI, Software Engineering Institute: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362

© 2022 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?

© 2022 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.

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

© 2022 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

© 2022 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.

Unintended and purposeful bias

- Bias can have purpose
- Bia can be helpful

Reduce unintended/unwanted and/or harmful bias.

© 2022 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 our algorithms...
Showing us the inequalities that have always been there."

Dr. Joy Buolamwini
Algorithmic Justice League
Movie: Coded Bias on Netflix

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

© 2022 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.

Transparency and accountability

Understand inherent bias and amount of variance in the data:

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

Provide evidence as appropriate – use context and knowledge of people using system.

© 2022 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

- Plan for AI system training and management of new data.
- Conduct continuous monitoring, evaluation, and audits 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

© 2022 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

Examine dynamic data and evaluate dynamic outcomes

- Is this the right data? What has changed?
- Is there enough 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

© 2022 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.

Implementing Responsible, Ethical, and Human-Centered AI

Moving Forward

© 2022 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.

How can teams harness the **power** of AI systems and design them to be **valuable** to humans?

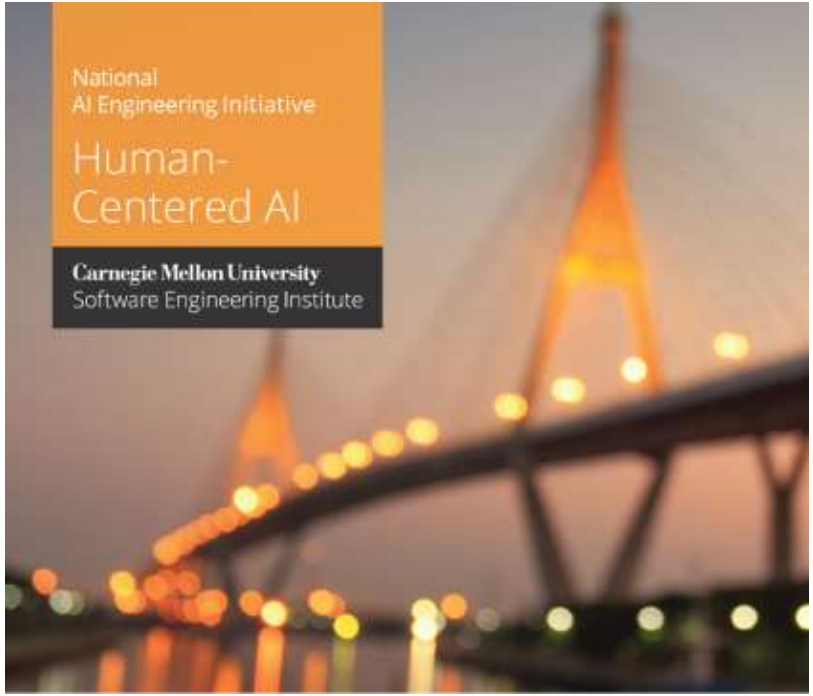
© 2022 Carnegie Mellon University

[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.

Advancing human-centered AI

Effective implementations minimize unintended consequences

- Understand complexity of context
- Design for human-machine teaming
- Engage in critical oversight

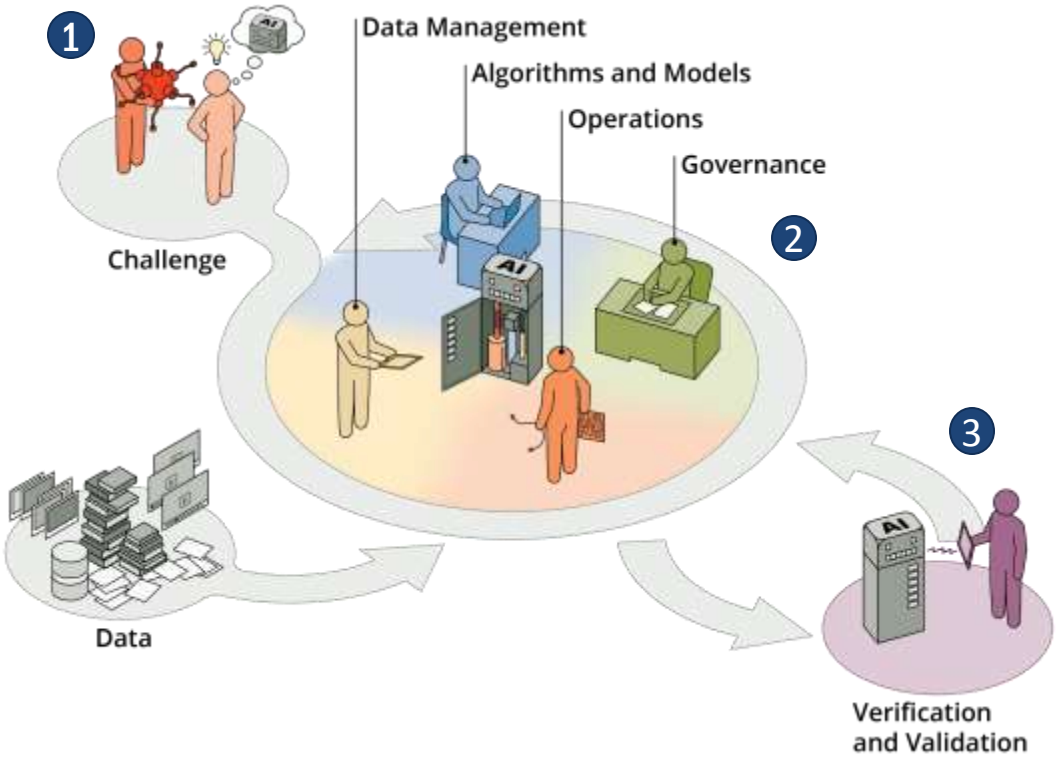


Human-Centered AI, Software Engineering Institute: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362

© 2022 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 Systems and Human-Machine Teaming



1. **Understand complexity of context:** focus on user needs
2. **Design for human-machine teaming:** align on technical ethics
3. **Engage in critical oversight:** Make AI interpretable, understandable, verifiable

© 2022 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.

Encourage deep conversations, speculation, imaginative thinking, and active curiosity.

© 2022 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.

Thank You

Alex Van Deusen
Design Researcher, CMU SEI
arvandeusen@sei.cmu.edu

Carol J. Smith
Sr. Research Scientist, CMU SEI
cjsmith@sei.cmu.edu

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.