BayCH
UX and AI Series, July 13, 2021

Beyond Interaction: Human-Machine Teaming

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

Twitter: @carologic @sei_etc

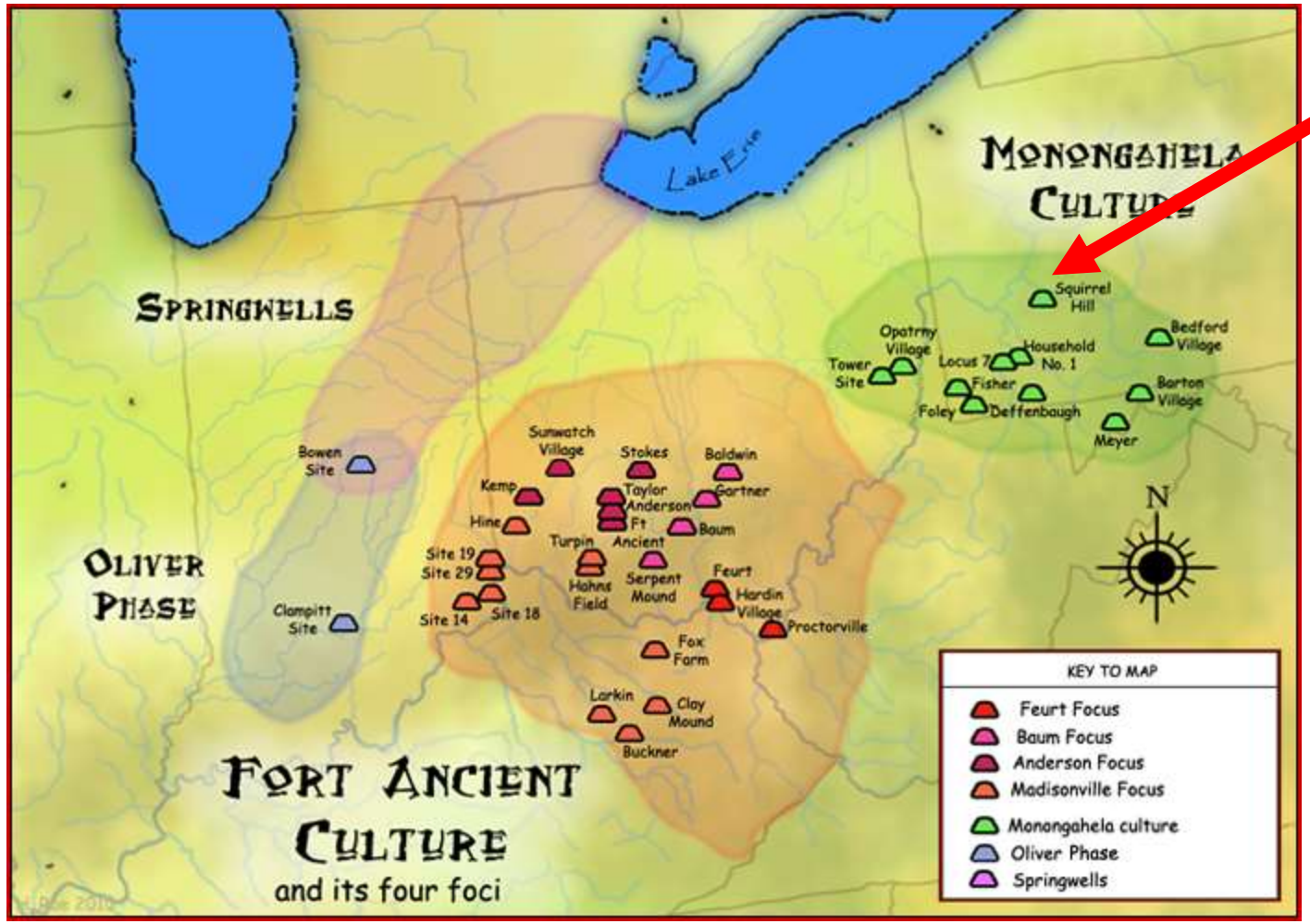
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

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.

DM21-0622

Acknowledgement: The Land I Speak On



Land of Monongahela, Adena and Hopewell Nations;
Seneca, Lenape and Shawnee lands;
Osage, Delaware and Iroquois lands.

Now known as Pittsburgh, PA, USA.

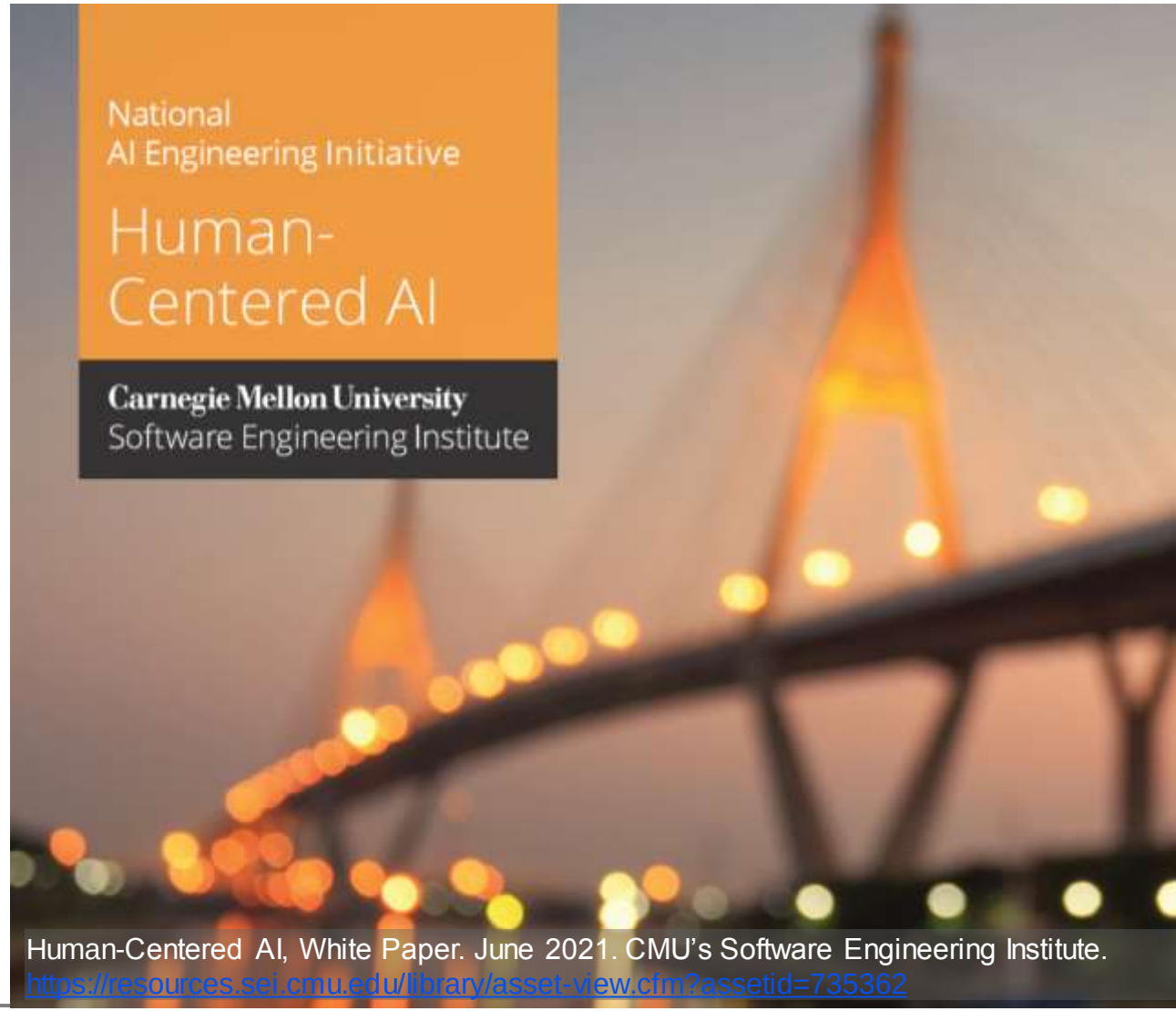
Map by Herb Roe via Wikipedia https://en.wikipedia.org/wiki/Monongahela_culture

Designed to work with, and for, people

Effective implementations

Minimize unintended consequences

1. Understand context of use
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 context of use

Human-Machine Teaming

Collaboration

- Human
- Machine (AI system)
- Information and interactions
- Exchanging information, potentially forming a relationship

Early, purposeful work

Is this an AI-friendly challenge?

For whom? What are their needs?

What kind of improvements are expected?

How will we know we've made improvements?

What are the benefits and risks?

Collaborative Activity/Event

- Time cycle - length of time interactions occur
- Length varies
  - Very short and hectic
  - Longer and iterative
  - Affects interactions

Clear communication, negotiation, and coordination required

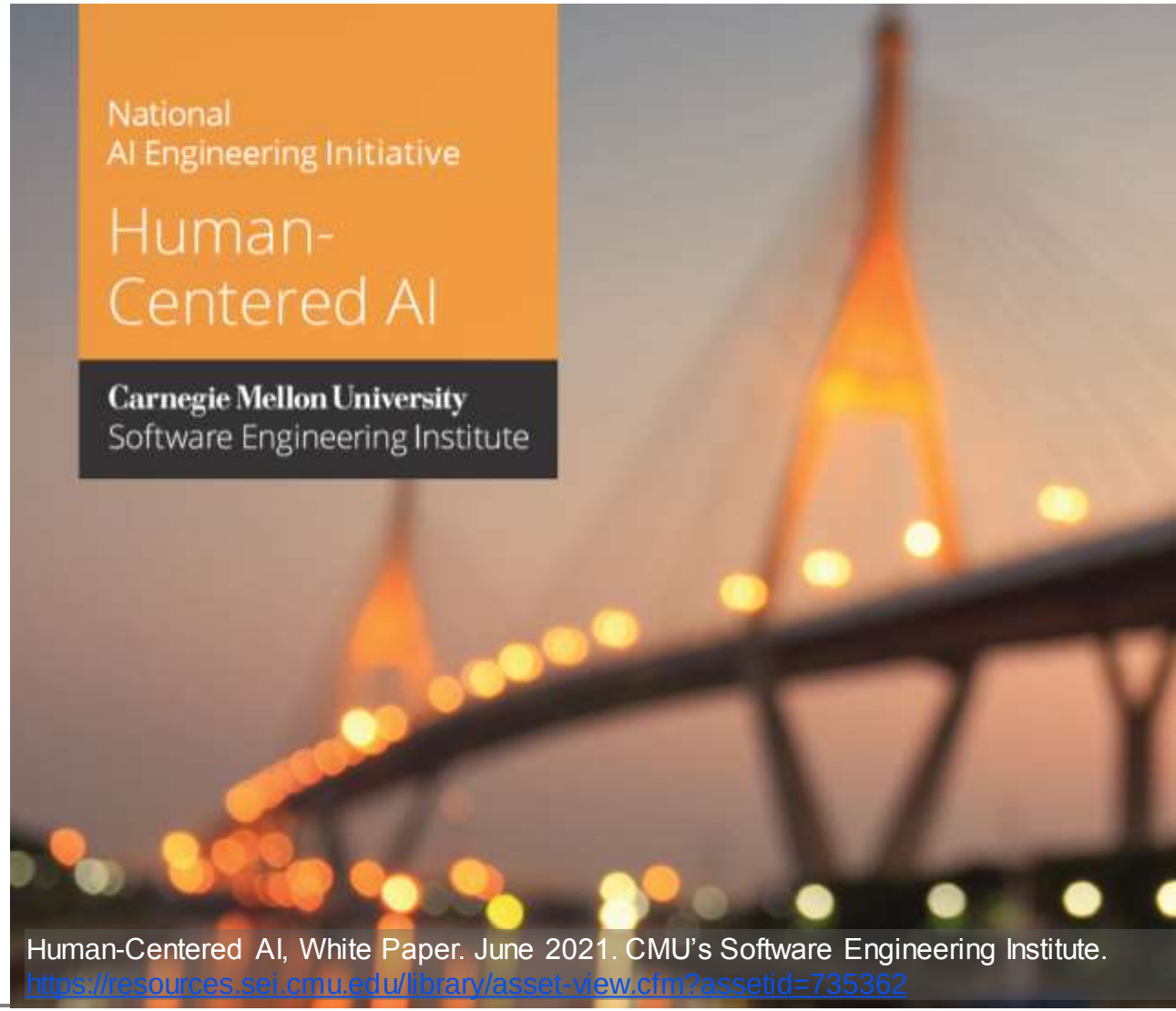
Context of use

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

Semi-autonomous vehicles

Potential factors

- Driver behavior
- Weather changes
- Street painting changes
- Change in desired route
- Highway vs. city driving
- Emergent situations

Image from Richard Marks on July 16, 2016 on Quora: https://www.quora.com/Why-does-the-UK-have-zig-zag-road-markings

Decision making for medical treatment

Potential factors

- How much information is already known?
- Stage of disease?
- Specifics of the patient's health, family situation, insurance status, etc.?
- What changes over time? New information?

Challenges

Getting time and budget for UX research

Using existing use patterns wisely

Demystifying AI to colleagues

Understanding what is different with regard to time cycles

Creating clear communication, negotiation, and coordination

Development of tools, processes, and practices

Design for human-machine teaming

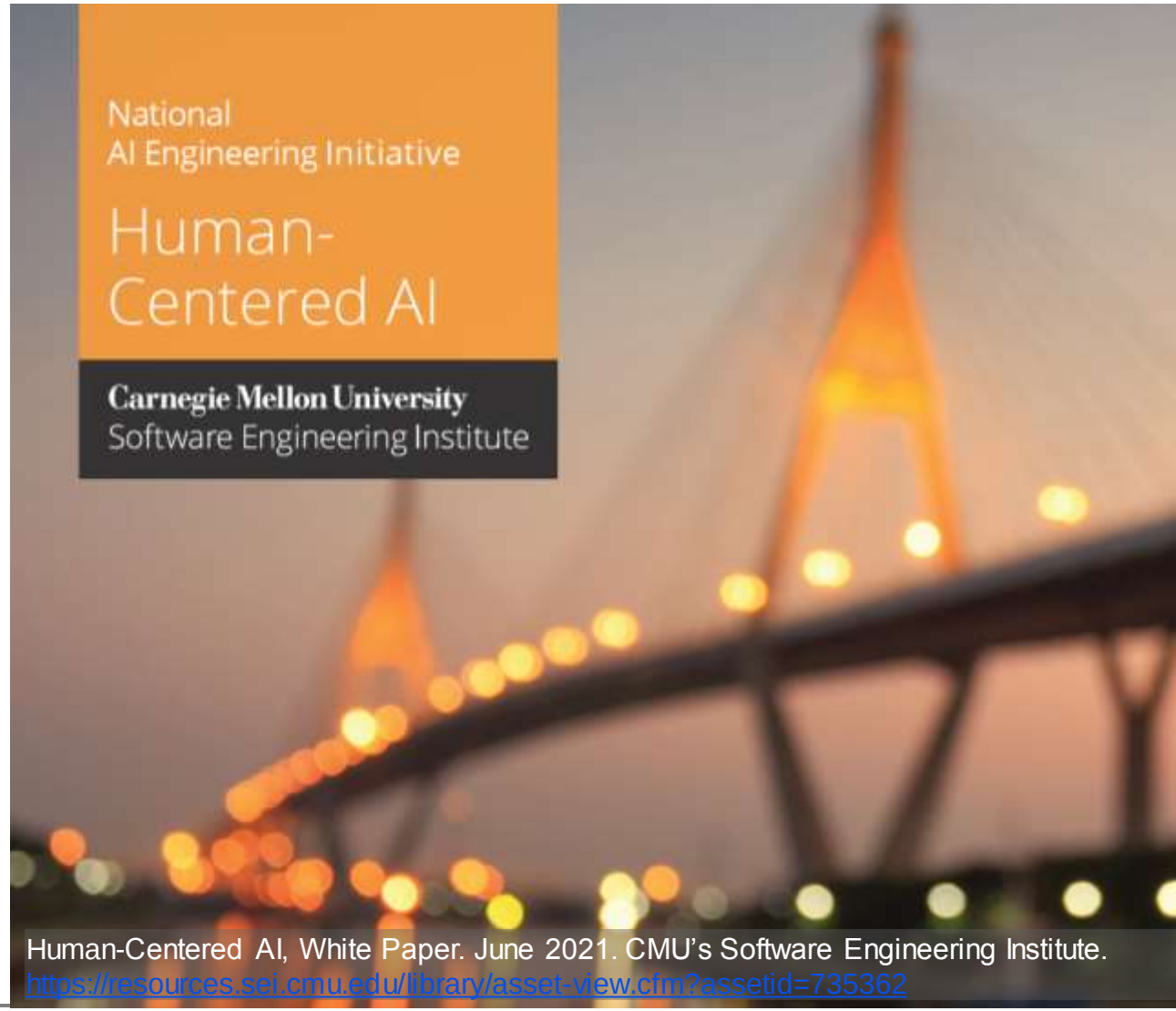
Interdependence

People interacting with and understanding systems

Gaining *appropriate* levels of trust

Design AI system to

- recognize boundaries and unfamiliar scenarios
- 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

AI is NOT sentient or unknowable

No one should use (or buy, or maintain), *important* systems that are described as indecipherable

Must have ability to control and monitor systems

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)
- Flip it to test it
- Implicit Association Test from Harvard University

Speculate about system misuse and abuse

- What are potential unintended/unwanted consequences?

More methods to "Outsmart Your Own Biases.": https://hbr.org/2015/05/outsmart-your-own-biases
Implicit Association Test (IAT): https://implicit.harvard.edu/implicit/takeatest.html

Reward team members for finding ethics bugs

Dr. Ayanna Howard
- on the Artificial Intelligence Podcast with Lex Fridman

Conversations for Understanding

UX Framework guides AI teams

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?

*"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 Checklist with Technical Ethics

- Bridges gap 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.

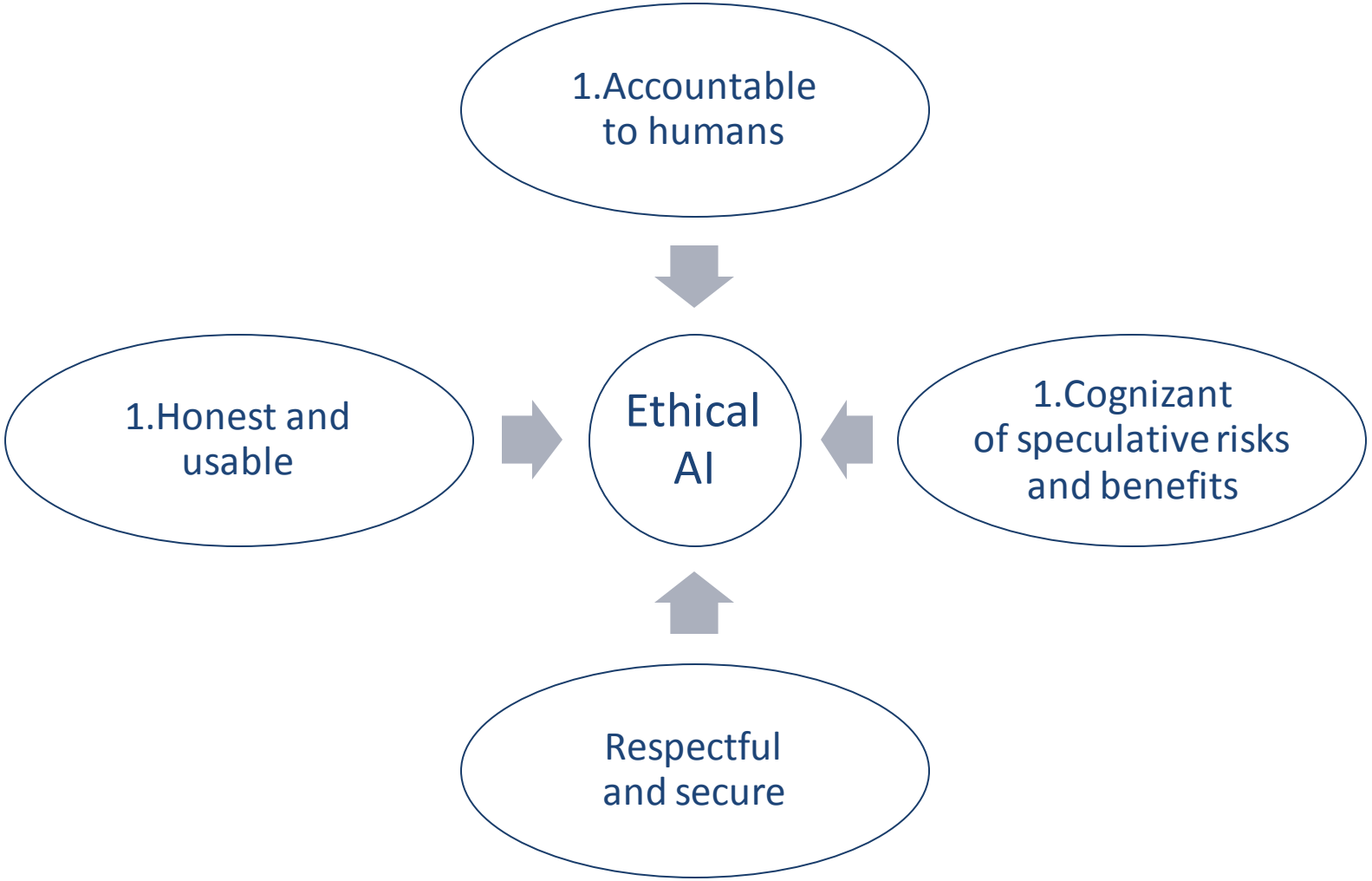
Team Signatures and Date

About the SEI

Contact Us

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

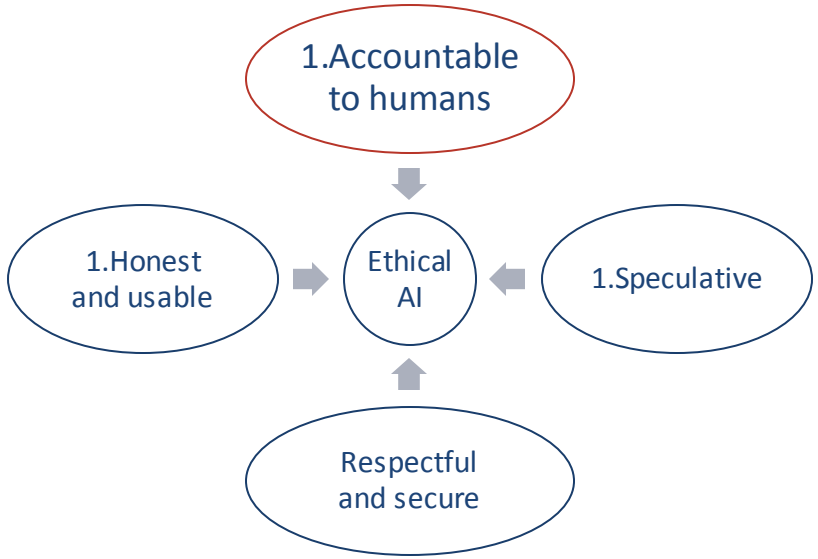
Accountable to Humans

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

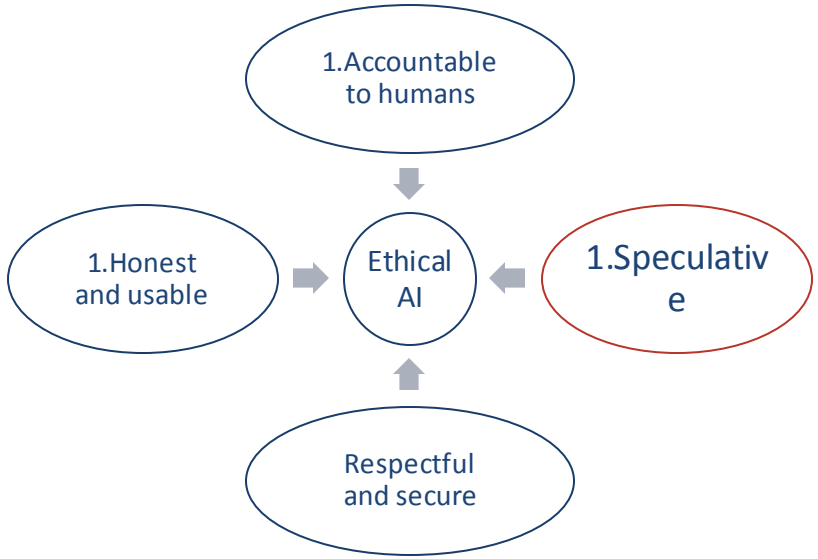
Feature added to enable RightStaff to turn off by itself

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

Cognizant of Speculative Risks and Benefits

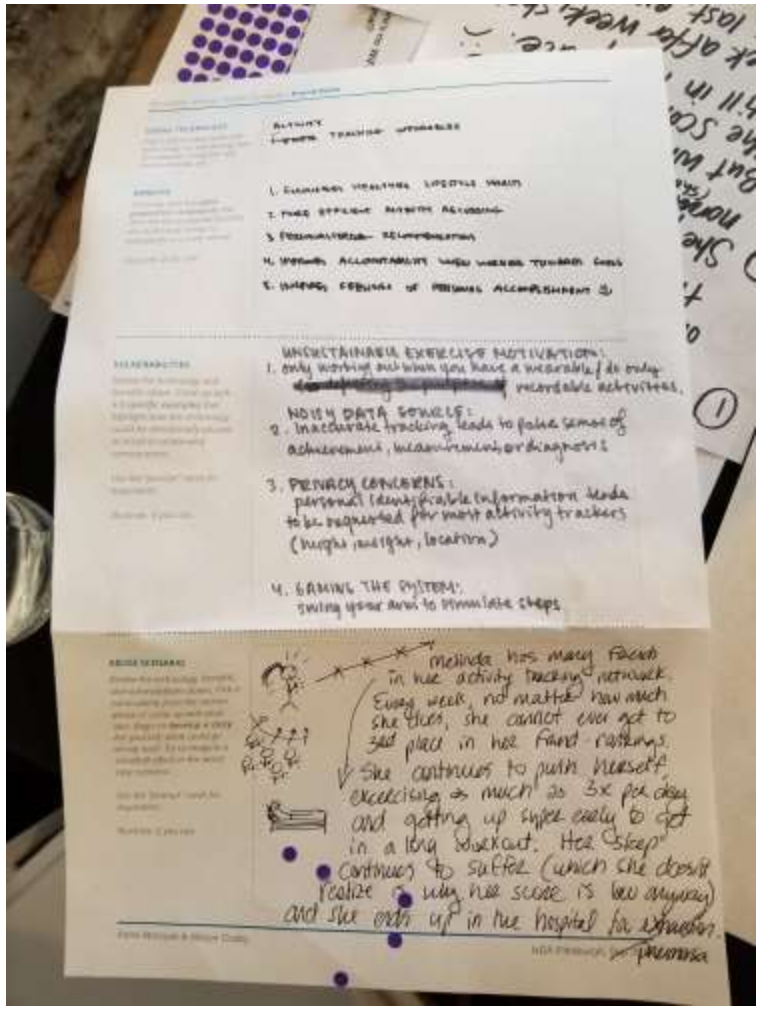
Identify full range of

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



Speculative: Conduct UX research - activate curiosity

- Speculate about misuse and abuse
- Potential severe abuse and consequences
- Perspective of people in frequently marginalized groups
- “Black Mirror” episodes



“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?

Speculative: Create communication & mitigation plans

Plan for unwanted consequences

Misuse and abuse of AI system

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

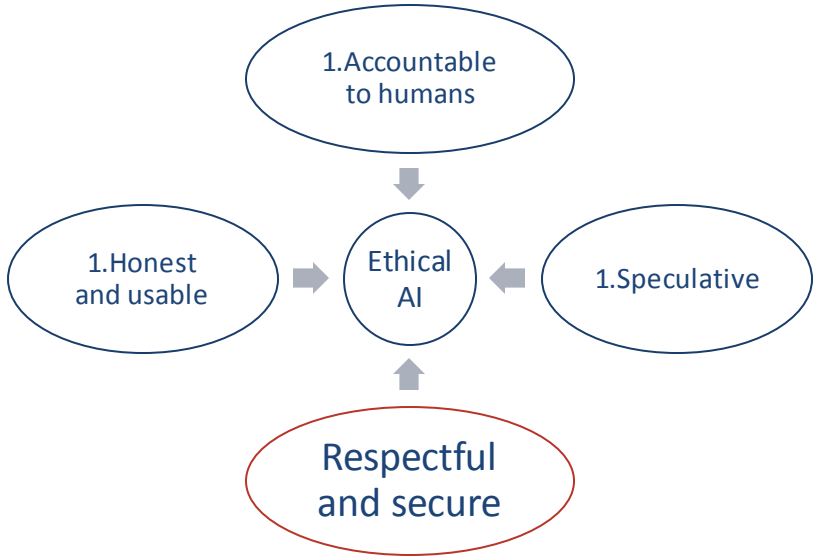
Respectful and Secure

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

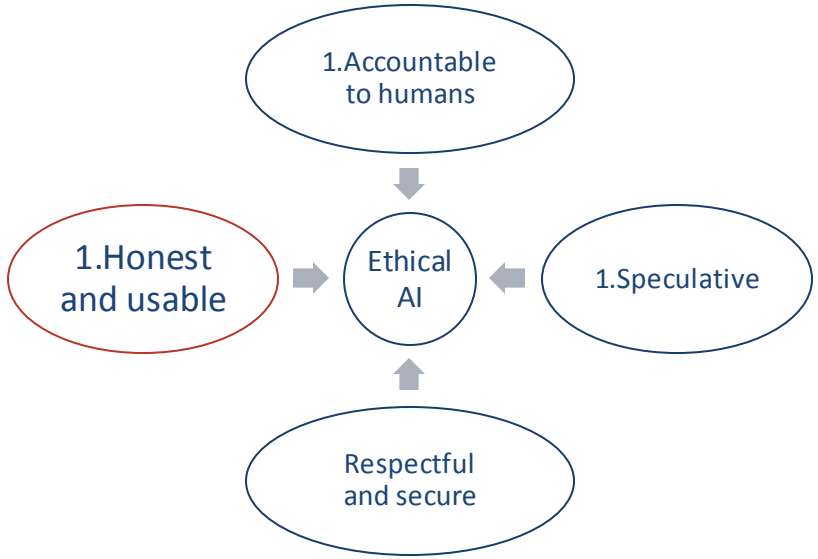
- Who has visibility to reasons for changing schedules?
- How is that information used?
- How is PII* of employees protected?

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

Honest and Usable

Value transparency with the goal of engendering trust

Explicitly state identity as an AI system



Fair: Remove 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)

Challenges

Need more speculative activities

Engaging people in this work is hard, and necessary (much like UX/HCI and accessibility)

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

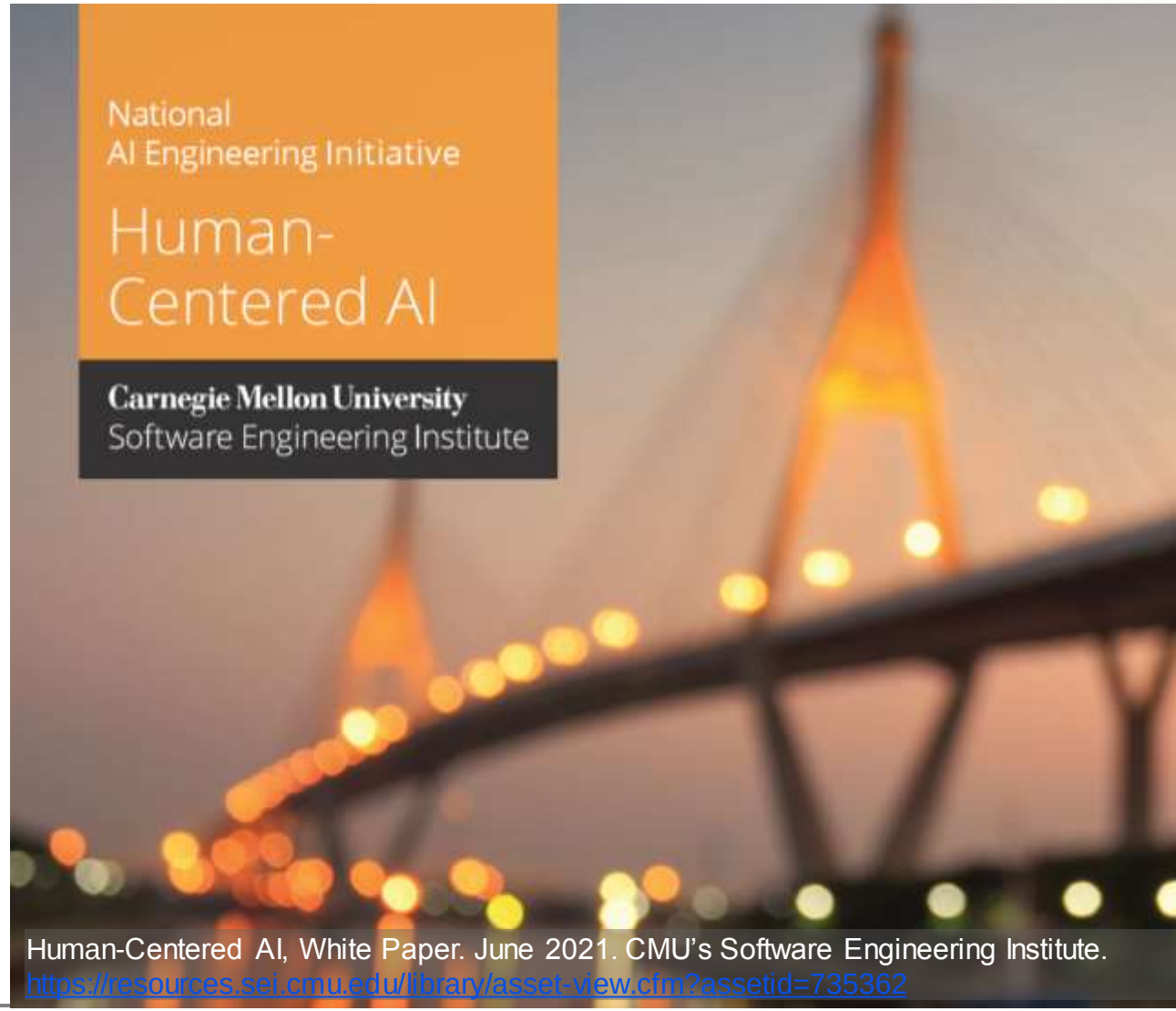
Methods, Mechanisms, and Mindsets

Engage in Critical Oversight

Continuous human oversight

Humans asking:

"What are we doing?
Why are we doing it,
and for whom?"



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

Risks are ever present

Proactively consider risks

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

AI should not be assumed to be “stable” – it is dynamic

Data transparency

Must understand data at a deep level

Provenance and creator's motivation

- What data included? Why?
- What not included? Why?

Use

- 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

How can Data be biased?

Goal: Select the right lawn care treatment. Save time.

Data: Multiple data source choices?

Selecting Data Source

Company A

- Primarily uses chemicals to treat lawns
- Data likely biased towards chemical use

Company B

- Only “all natural” treatments
- Data likely biased against chemical use

Selecting Data Source

Company A

- Primarily uses chemicals to treat lawns
- Data likely biased towards chemical use

Company B

- Only “all natural” treatments
- Data likely biased against chemical use

Neither are wrong.
Both are biased.

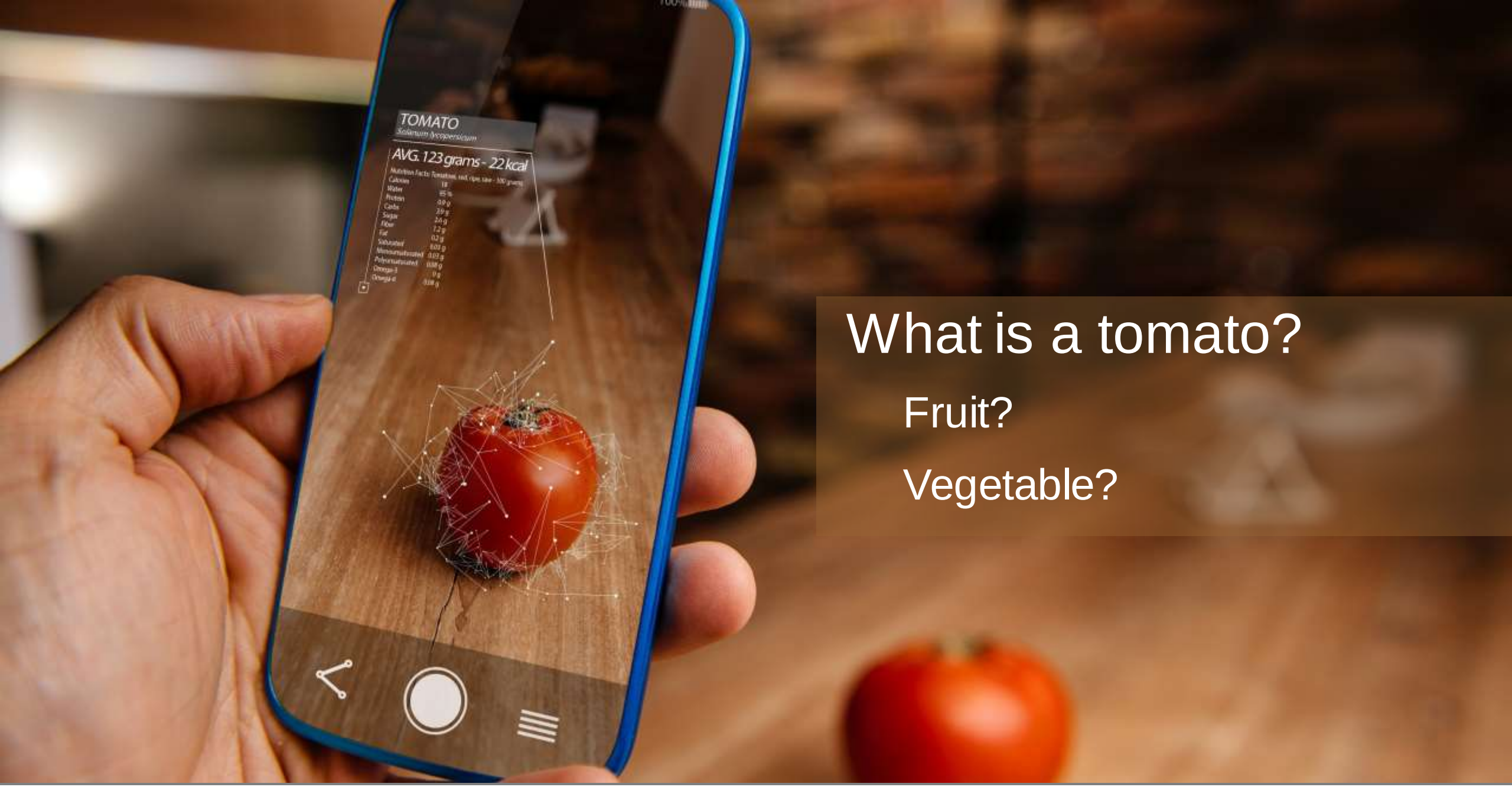
Bias in data, algorithm selection, and training

Unintended and purposeful bias

Misuse and abuse of the system

Understand inherent bias and amount of variance.

Document the data's motivation, composition, collection process, recommended uses, etc. for transparency and accountability.



What is a tomato?

Fruit?

Vegetable?

Bias in Image Recognition

Training data

Data encountered

Only know what taught

Training data

Unrepresentative
or incomplete training data

Data encountered

Unlikely to recognize

Joy Buolamwini, Algorithmic Justice League

Coded gaze

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

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

Categories of Harm

Use these categories of harm to evaluate how a product, service, or technology could cause harm.

CATEGORIES OF HARM	DEFINITION
FINANCIAL	Negative impact on finances, property, or other resources
HEALTH	Negative impact on mental, emotional, or physical health
TIME	Inefficient or unproductive activities, processes, or systems
FAIRNESS/EQUITY	Perpetuating or facilitating prejudice, bias and/or unfairness
SAFETY	Physical and/or emotional wellbeing compromised by fear, danger, or uncertainty
PRIVACY	Lack of control over personal information
MISINFORMATION	The creation, spread and/or amplification of false or inaccurate information intended to deceive
CONTROL	Inability to freely direct information, activities, or systems
TRANSPARENCY	Lack of disclosure of information, activities, or systems

Leaders must establish psychological safety

Datasheets for Datasets

Transparency and clarity

- Creator's motivation
- Composition
- Collection
- Preprocessing / Cleaning / Labeling
- Uses
- Distribution
- Maintenance

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

Auditing

Probe with hypothetical cases

Checks for bias, brittleness or potential distribution shift

Access history of system operation and usage*

*Consider ethical principles when determining what data needs to be collected.

Challenge: Shift work

Examining dynamic data and evaluating dynamic outcomes

- Is this the right data? What has changed?
- Did the system respond appropriately given the situation?
- Was the AI an effective collaborator?
- Was the human willing to give appropriate trust?

We must work to define standard methods and processes for evaluating system outcomes

AI has great potential, develop with caution

Future AI's may be trusted to substitute human cognition and abilities.

Humans must continue to be responsible for situations that involve a person's:

- Life (the use of force)
- Quality of life
- Health
- Reputation

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

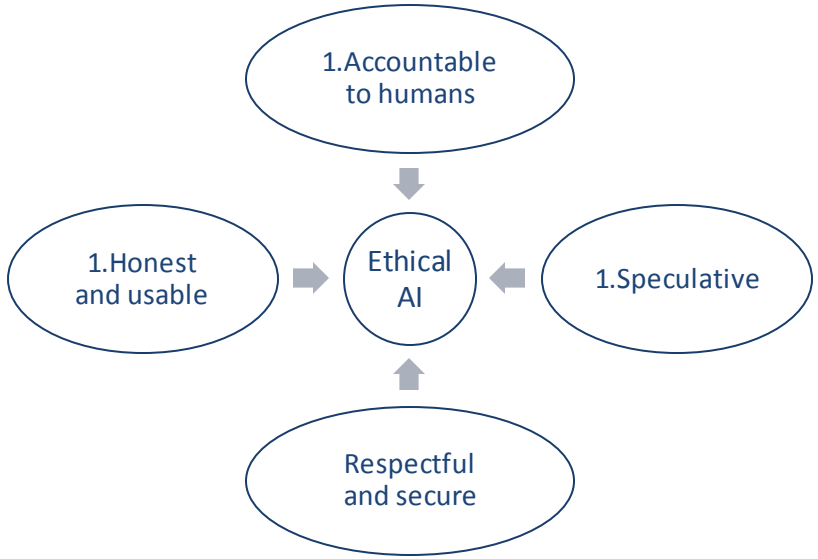
We aren’t perfect, AI won’t be perfect

Empower diverse teams, inclusive environments

Adopt technical ethics

Encourage deep conversations

Activate curiosity; be speculative; imaginative



Evangelize for human values

Ethical. Responsible. Transparent. Fair.

Carol J. Smith

Twitter: @carologic

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

CMU's Software Engineering Institute,
Emerging Technology Center

Twitter: @sei_etc

Explainable AI

DARPA coined term for the experience of ML engineers, developers, and others dealing with the inner workings (algorithms, etc.). This work is to ensure that what the system is doing and why is understandable to the people building and maintaining it.

The two experiences are completely different, and in both cases, work needs to be done quite early in the development process to ensure that the system is . This work is hard and necessary.

People working in AI have a responsibility to ensure that people have the ability to control and monitor any system we build.

No one should use (or buy, or maintain), important systems that are described as indecipherable* or that they do not have the ability to control and monitor.

Coalesce on Shared Set of Technology Ethics

1. Well-being
2. Respect for autonomy
3. Protection of privacy and intimacy
4. Solidarity
5. Democratic participation
6. Equity
7. Diversity inclusion
8. Prudence
9. Responsibility
10. Sustainable development

Montréal Declaration for a responsible development of artificial intelligence.
https://www.montrealdeclaration-responsibleai.com/the-declaration