CMU XHacks

Ethics and Trust for Emerging Technologies

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

Twitter: @carologic @SEI_CMU_AI

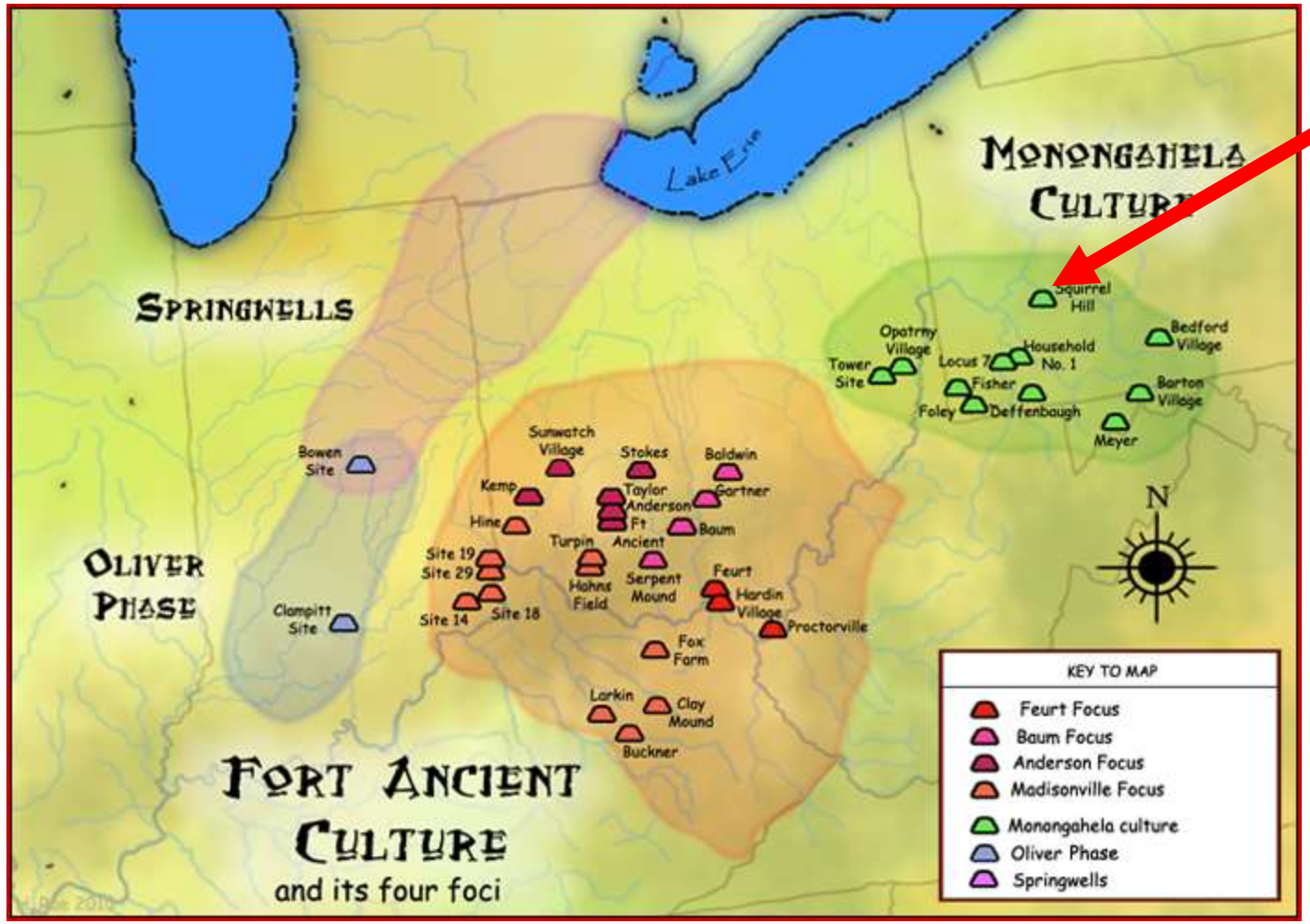
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

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.

DM22-0292

Acknowledge This Land



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

Land of Monongahela, Adena and Hopewell Nations;

Seneca, Lenape and Shawnee lands;

Osage, Delaware and Iroquois lands.

Now known as Pittsburgh, PA, USA.

Emerging technology

Great potential - develop with caution

abc NEWS

Ring security camera hacks see homeowners subjected to racial abuse, ransom demands

A spate of incidents has seen homeowners in four states fall victim to hackers.

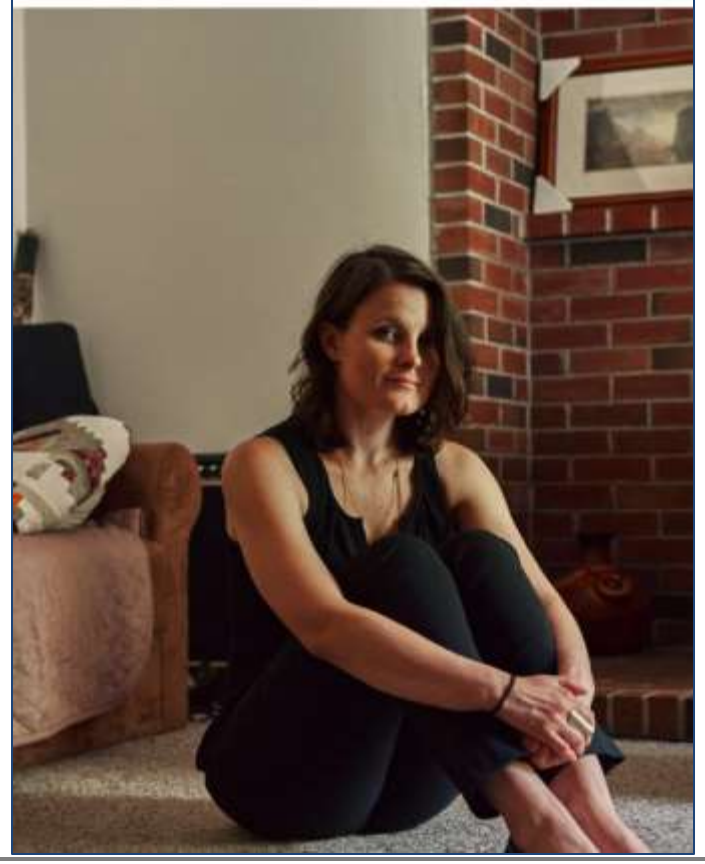
By Mark Hanrahan
December 12, 2019, 9:56 PM • 7 min read

Ring camera systems being hacked

Multiple U.S. families have reported incidents of Ring camera systems being hacked in recent days.

The New York Times

Thermostats, Locks and Lights: Digital Tools of Domestic Abuse



Responsible, Intentional Design

Early, purposeful work

In addition to the usual UX work

- How will the system partner with people? Compliment?
- What are the obvious risks?
- How might these systems be misused/abused?

Ethics

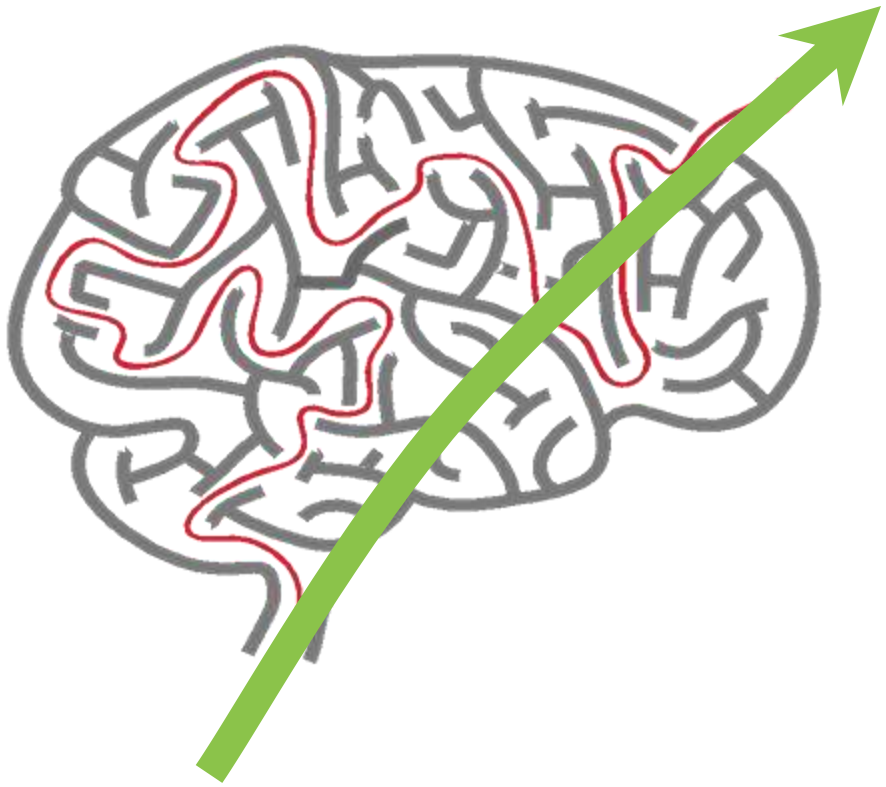
Based on well-founded standards of right and wrong

Standard of expected behavior that guides the correct course of action

What impact does my work have?

What is Ethics? By Manuel Velasquez, Claire Andre, Thomas Shanks, S.J., and Michael J. Meyer. Markkula Center for Applied Ethics
https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/what-is-ethics/

To be biased, is to be human



Bias are shortcuts, to avoid risk and simplify problems.

Not inherently bad, may be misapplied

Implicit = invisible

Not necessarily in sync with our conscious beliefs

Can be managed and changed

All systems have some form of bias

Data is collected/curated,
and systems are created
by humans for a purpose.

Complete objectivity is misleading.

Bias can have purpose and can be helpful.

Our Goal: **Reduce unintended and/or harmful bias.**

Bias in Image Recognition

Training data

Use case courtesy of Dr. Eric Heim, CMU SEI
https://resources.sei.cmu.edu/library/author.cfm?authorid=542374

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

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

High value in diverse teams

Diverse teams

- focus more on facts
- process facts more carefully
- are more innovative

"…become more aware
of their own potential biases"

Photo by Christina @ wocintechchat.com on Unsplash
https://unsplash.com/@wocintechchat?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText

David Rock, Heidi Grant. 2019. Why Diverse Teams Are Smarter. *Harvard Business Review*. November 4, 2019. https://hbr.org/2016/11/why-diverse-teams-are-smarter

Great minds think different

- How can we best identify the biases within ourselves and our teams?
- What tools and processes to control for bias can we introduce into our processes?
- Where do biases emerge in our work?

Diverse, inclusive leaders

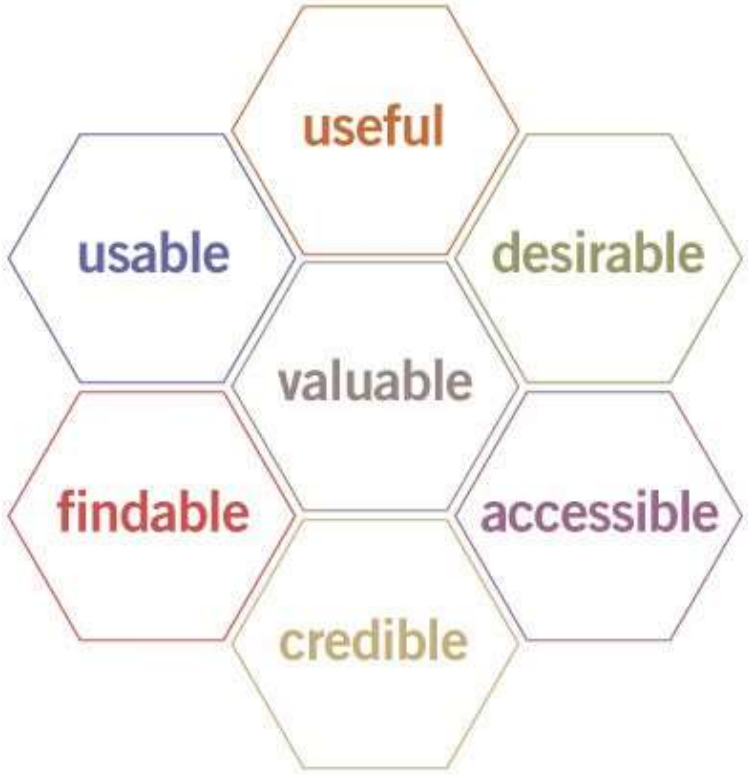
Diverse, Multi-Disciplinary Teams

Shared Tech Ethics

UX Framework

Implementing Ethics

How do we make tech human-centered?



User Experience Honeycomb
Peter Morville, et al.

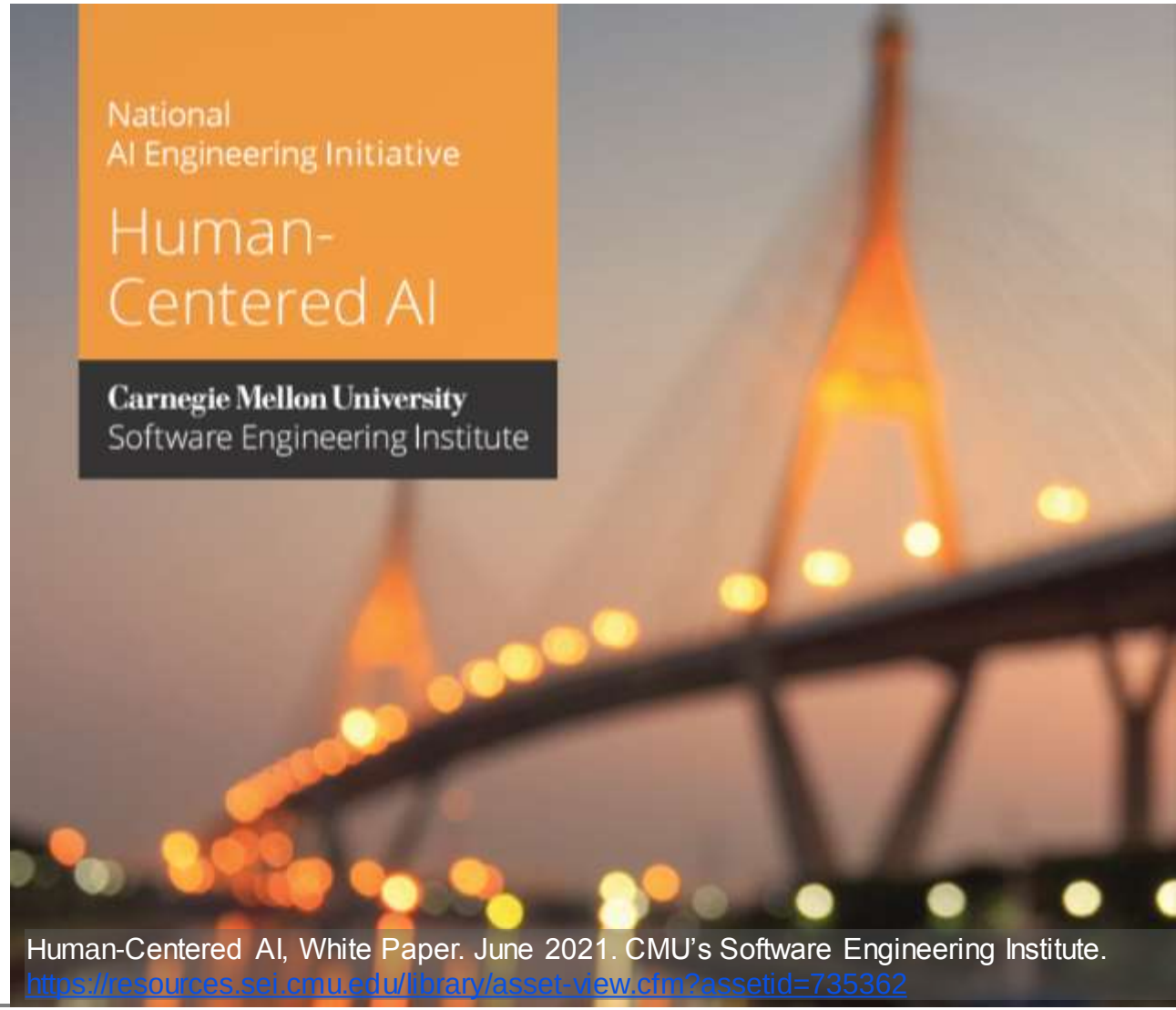
Human-Centered Technology

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

Change is constant

What changes across time cycles?

Length of interactions

- Short and hectic
- Longer, cyclical - iterative

Collaboration requires clear

- communication,
- negotiation, and
- coordination.

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

Context

User research to understand

- Complexity (environmental, human and information)
- Effect of change
- Determine how design responds to change
- Overall changes, over time and experience

Trust

Trust is complex, transient, and personal.

Trust is a person using evidence to determine risk.

For complex systems, when a person has enough confidence in positive outcomes, they give control of something significant, to the system.

Trust is a continuum

Calibrated based on the context and the available evidence of the system's capability and integrity.

Distrust
Trust falling short of system capabilities - may lead to disuse.

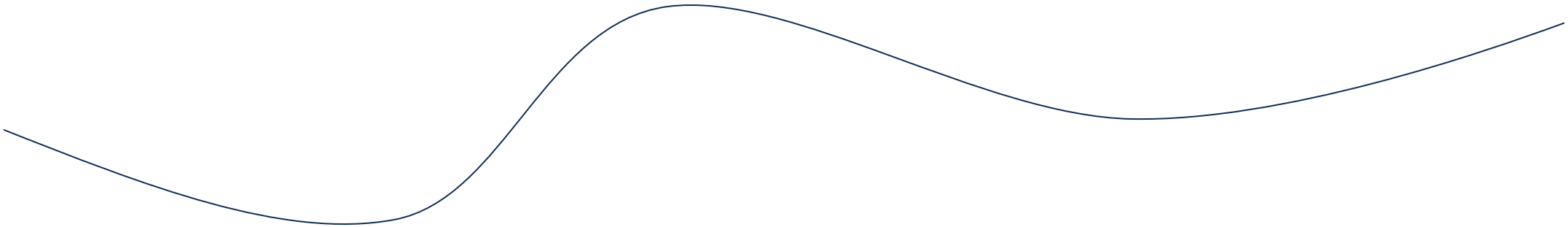
Calibrated Trust
Trust matches system capabilities leading to appropriate use.

Over Trust
Trust exceeding system capabilities - may lead to misuse

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

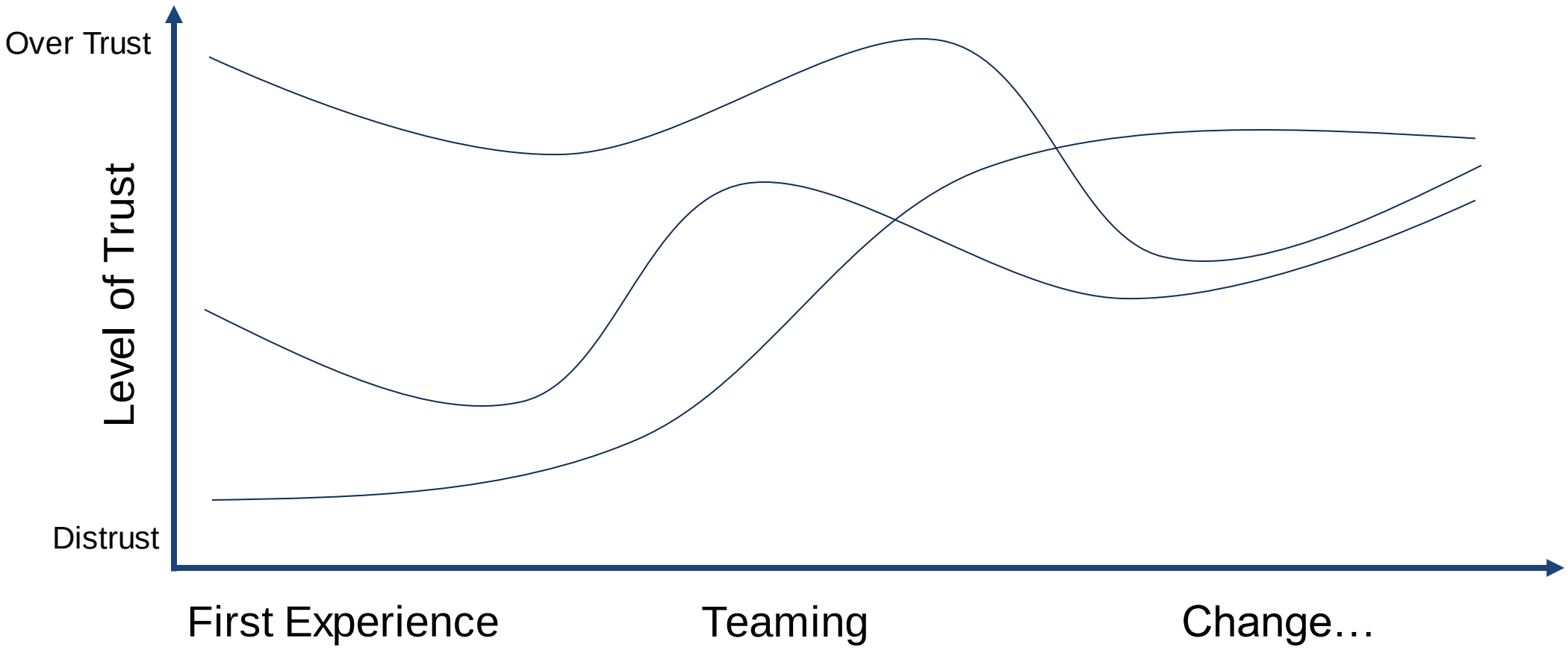
Calibrated trust is contextual

As the context changes, and/or as confidence in the AI system's capability and integrity change, there will be a corresponding change in the person's calibrated trust.



Building on 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

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

Changes increase or decrease trust

Event-Driven

- Response to an interaction, transaction, service, or event

Time-Driven

- Response to periodic evidence (observations, recommendations)
- Lack of evidence can decay trust

Jia Guo and Ing-Ray Chen. 2015. A Classification of Trust Computation Models for Service-Oriented Internet of Things Systems. 2015 IEEE International Conference on Services Computing (2015), 324-331. DOI: https://doi.org/10.1109/SCC.2015.52
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

Speculate and design for the worst case

Be speculative about the worst case

- Don’t assume only average cases
- Probabilities about what will happen in the future can't be verified
- Don’t require unsupportable risk assessments.

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

Conversations for Understanding

- What do we value?
- Who could be hurt?
  - Frequently marginalized groups
  - Tthose that are “unlucky”
- What lines won’t our system 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?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText

New uncomfortable work

“*Be uncomfortable*”

- Laura Kalbag

Ethical design is not superficial.

Coalesce on shared technology ethics

- Harmonize cultural variations
- Explicit permission to consider and question breadth of implications
- Pair with checklists to **prompt conversations**

Association for Computing Machinery

AINOW INSTITUTE

Microsoft

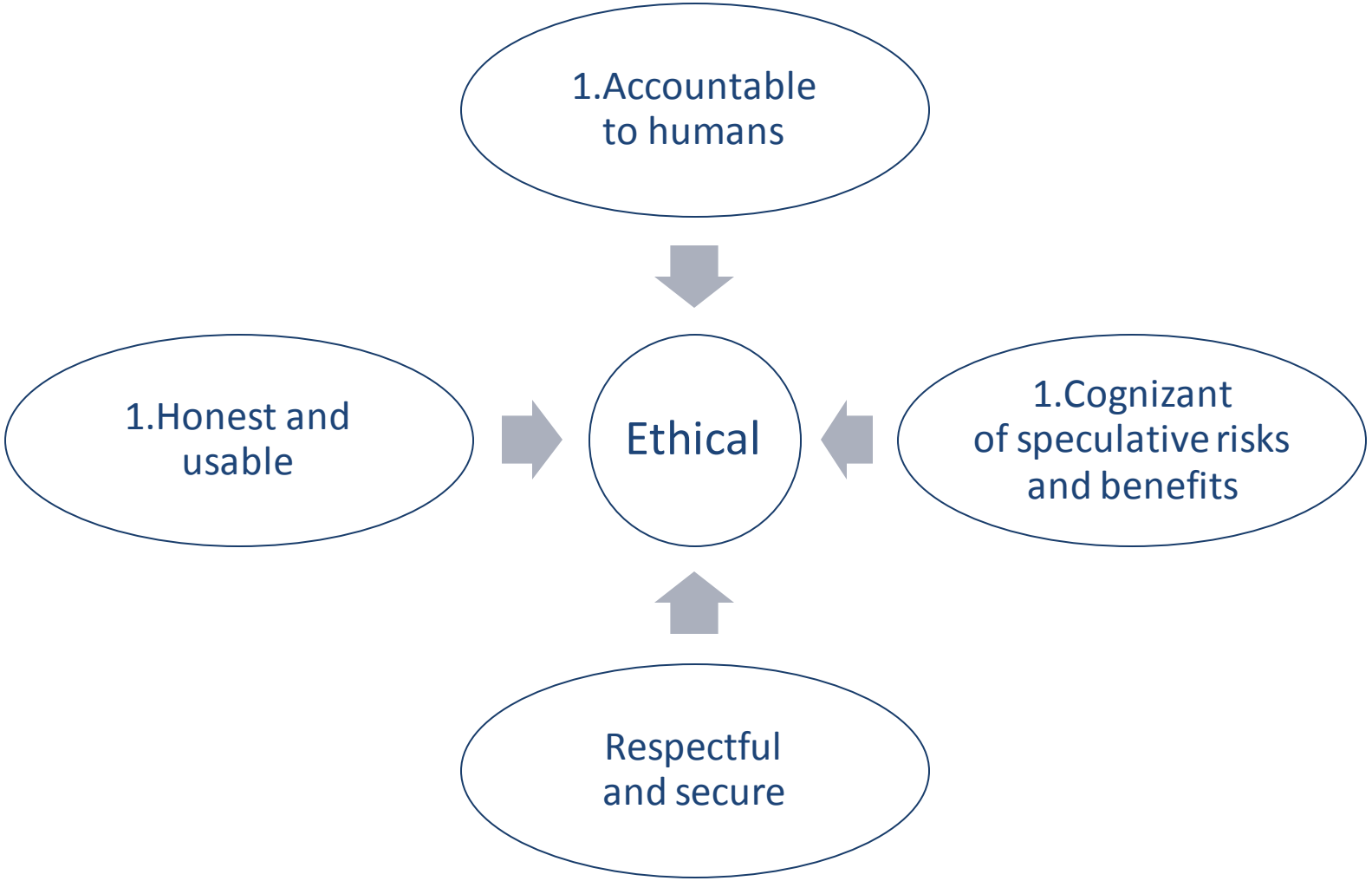
Google

< >
Montréal Declaration
Responsible AI_
</ >
An initiative of Université de Montréal

ETHICAL OS

U.S. DEPT OF DEFENSE

Prompt conversations – UX Framework



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

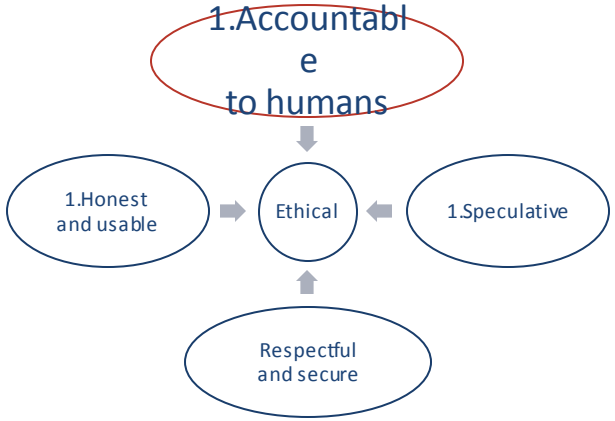
Accountable to Humans

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

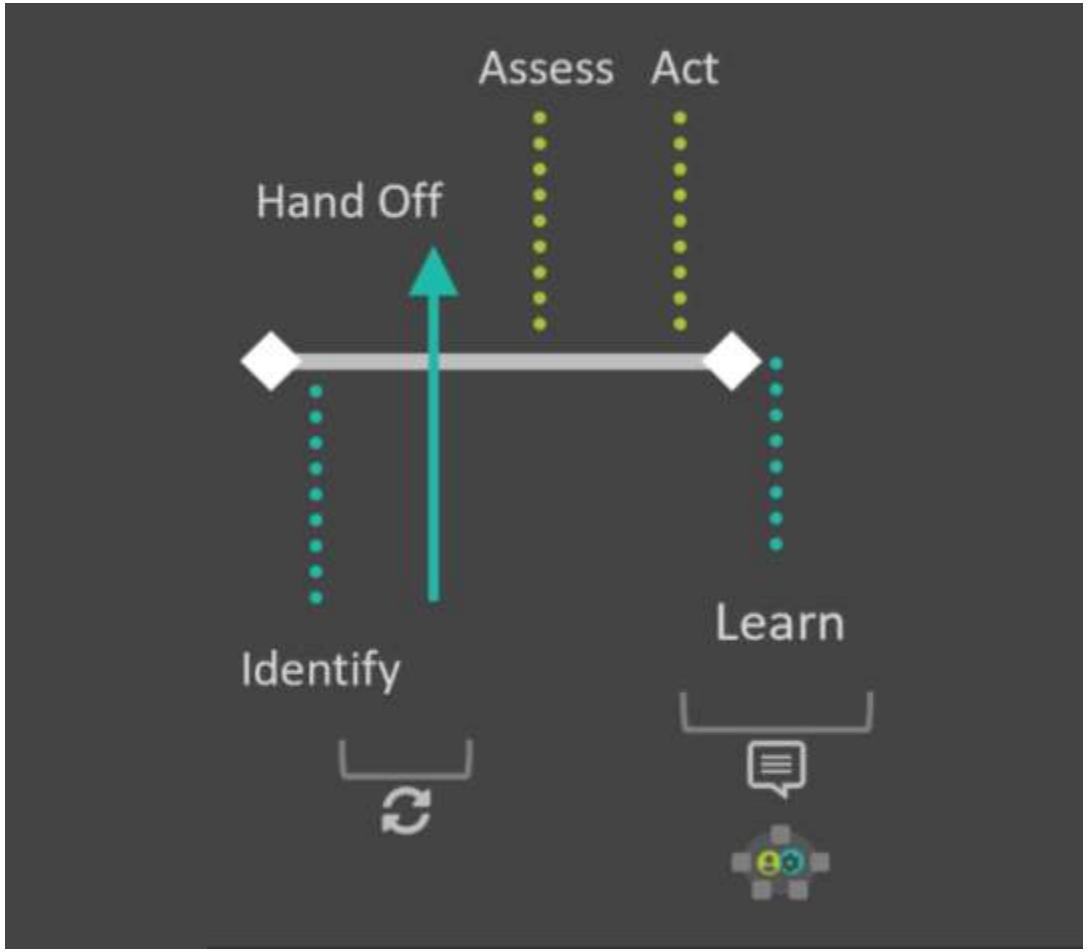
https://www.ted.com/talks/grady_booch_don_t_fear_superintelligence

Significant decisions

Significant decisions made by the system

- explained
- able to be overridden
- appealable and reversible

Responsibilities explicitly defined between people and systems

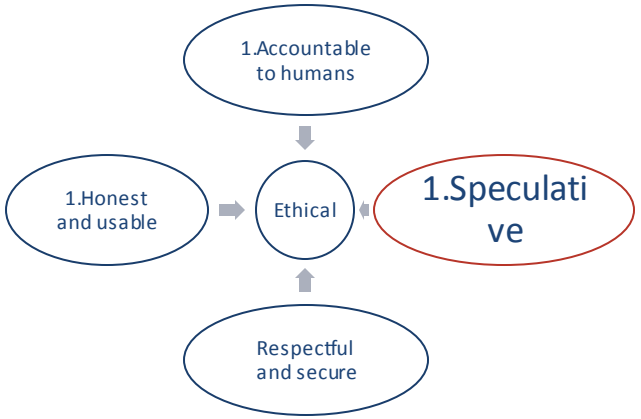


How IAs Can Shape the Future of Human-AI Collaboration. Carol Smith and Duane Degler.
Presented on April 28-30, 2021 at the Information Architecture Conference (IAC21)

Cognizant of Speculative Risks and Benefits

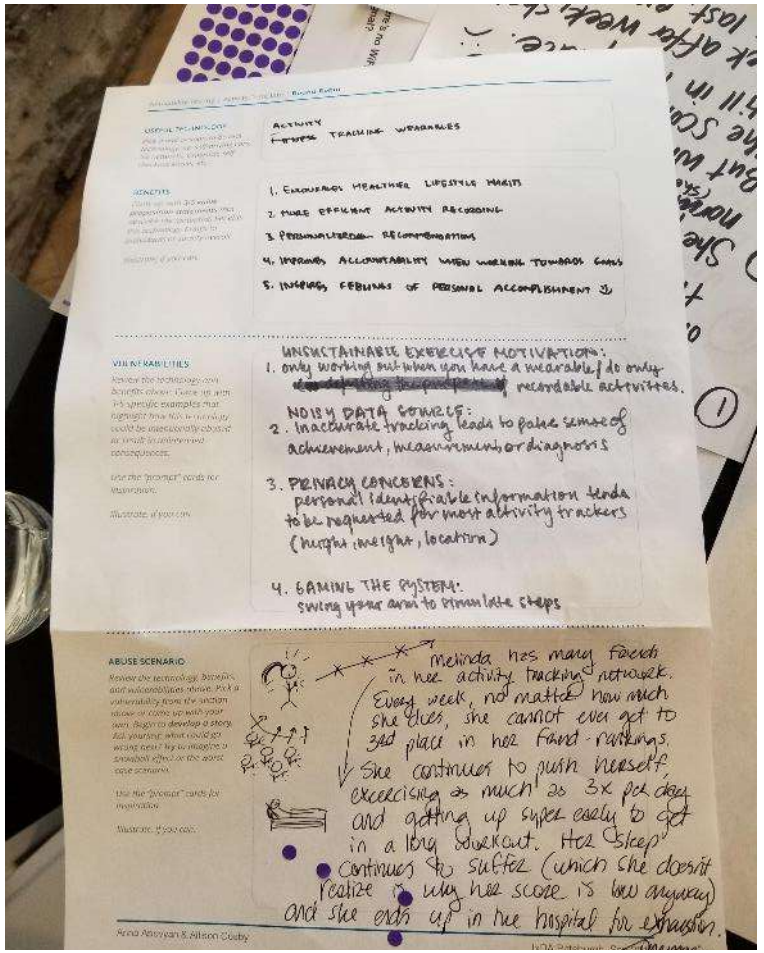
Identify full range of

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



Conduct UX research - activate curiosity

- Speculate about misuse and abuse – abusability testing
- Potential severe abuse and consequences
- Perspective of people in frequently marginalized groups



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

Create communication & mitigation plans

Plan for unwanted consequences

Misuse and abuse of system

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

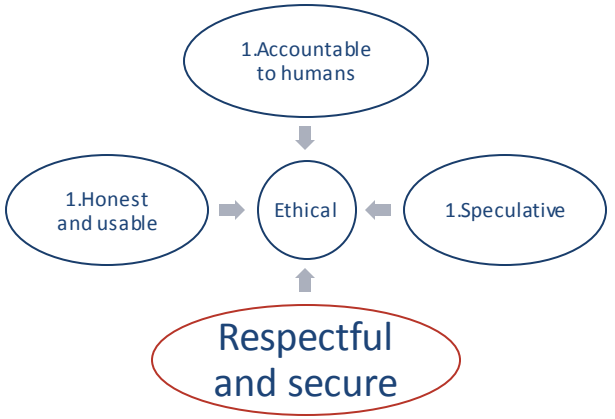
Respectful and Secure

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

Respect privacy and data rights

Make system robust, valid and reliable

Provide understandable security



Honest and Usable

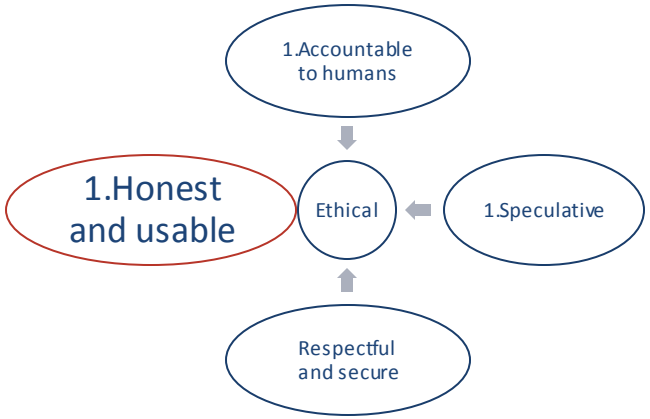
Value transparency with the goal of engendering trust

Smart speakers explicitly state identity as an AI system

Remove unwanted bias in data

Show awareness of known and desirable bias

Acknowledge issues



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

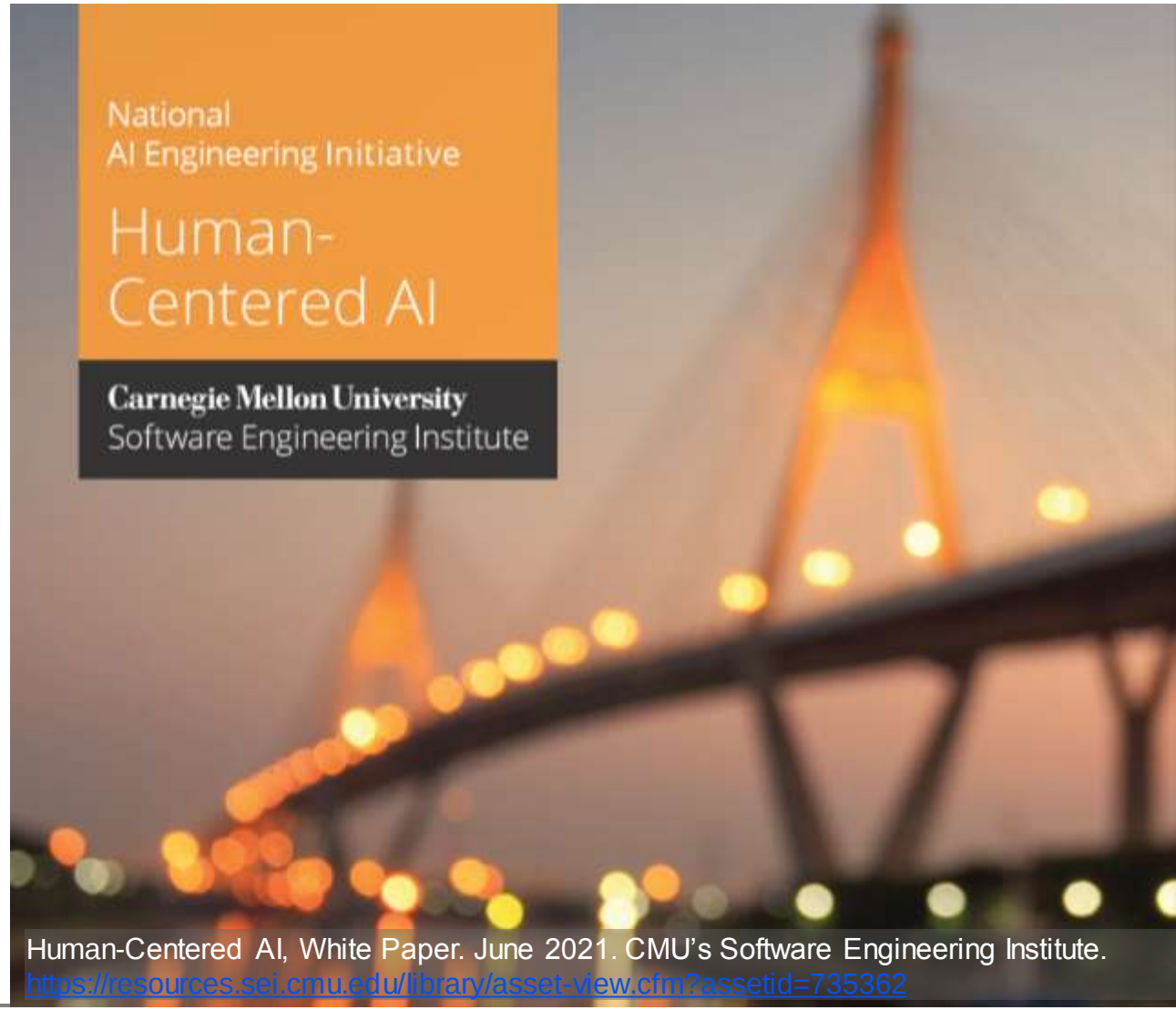
Engage in critical oversight

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

Continuous human oversight

Identify risks of bias, misuse, abuse, and unintended consequences

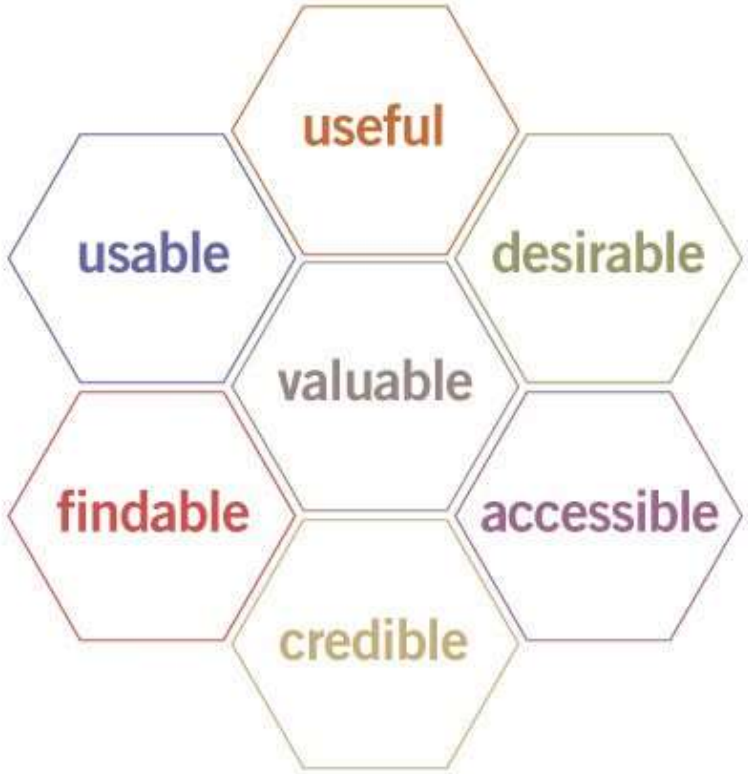
Proactively consider risks



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

Leaders must 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

Human-Centered AI

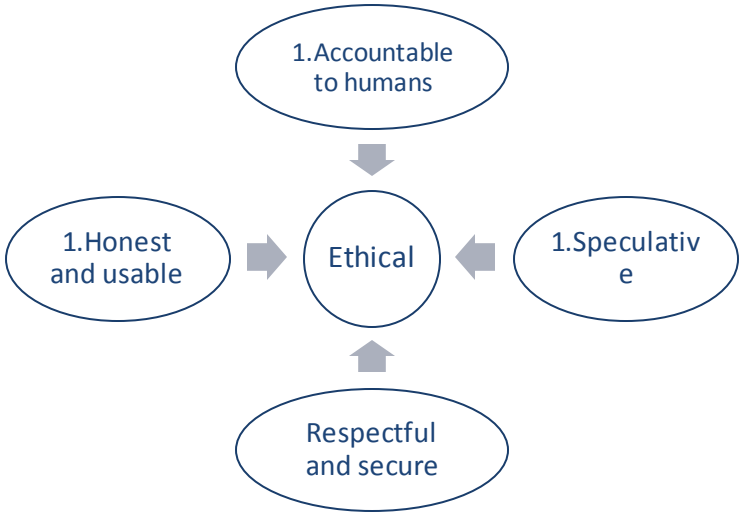
Activate curiosity. Be speculative. Imaginative.

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

Empower diverse teams, inclusive environments

Adopt technical ethics

Encourage deep conversations



It’s Your Turn!

Great user experiences match with people’s needs and are:

- Equitable
- Sustainable
- Accessible

Create innovative tech solutions!

Carol J. Smith

Twitter: @carologic

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

CMU Software Engineering Institute,
AI Division

Twitter: @SEI_CMU_AI

Additional Resources

Montréal Declaration for a responsible development of artificial intelligence

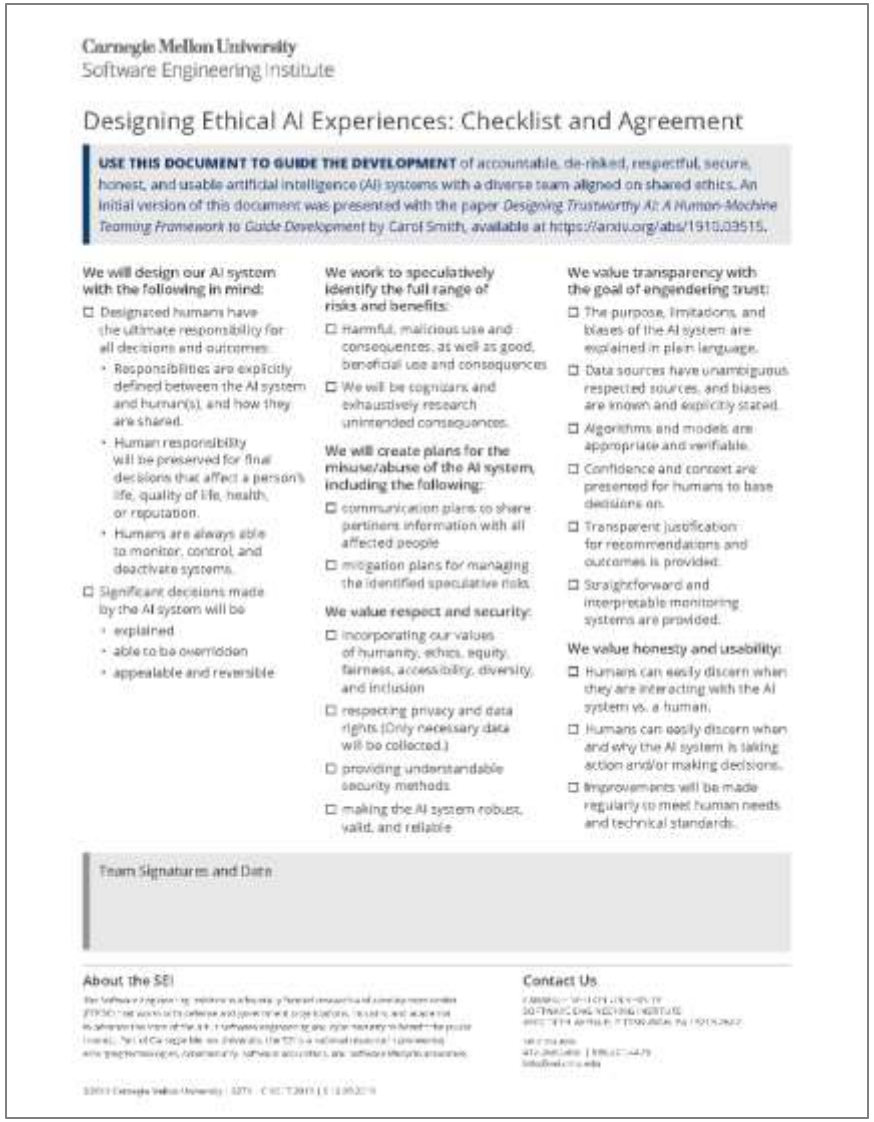
1. Well-being
2. Respect for autonomy
3. Protection of privacy and intimacy
4. Solidarity
5. Democratic participation
6. Equity
7. Diversity inclusion
8. Prudence
9. Responsibility
10. Sustainable development

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

Checklists to 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

Automation bias

Propensity for humans to favor
suggestions from automated decision-making systems

and to **ignore contradictory information made without automation**,

even if it is correct.

Mary Cummings. 2004. Automation Bias in Intelligent Time Critical Decision Support Systems. AIAA 2004-6313. AIAA 1st Intelligent Systems Technical Conference. (September 2004). DOI: https://doi.org/10.2514/6.2004-6313

DEFENSE INNOVATION UNIT

ARTIFICIAL INTELLIGENCE PORTFOLIO

Responsible AI Guidelines

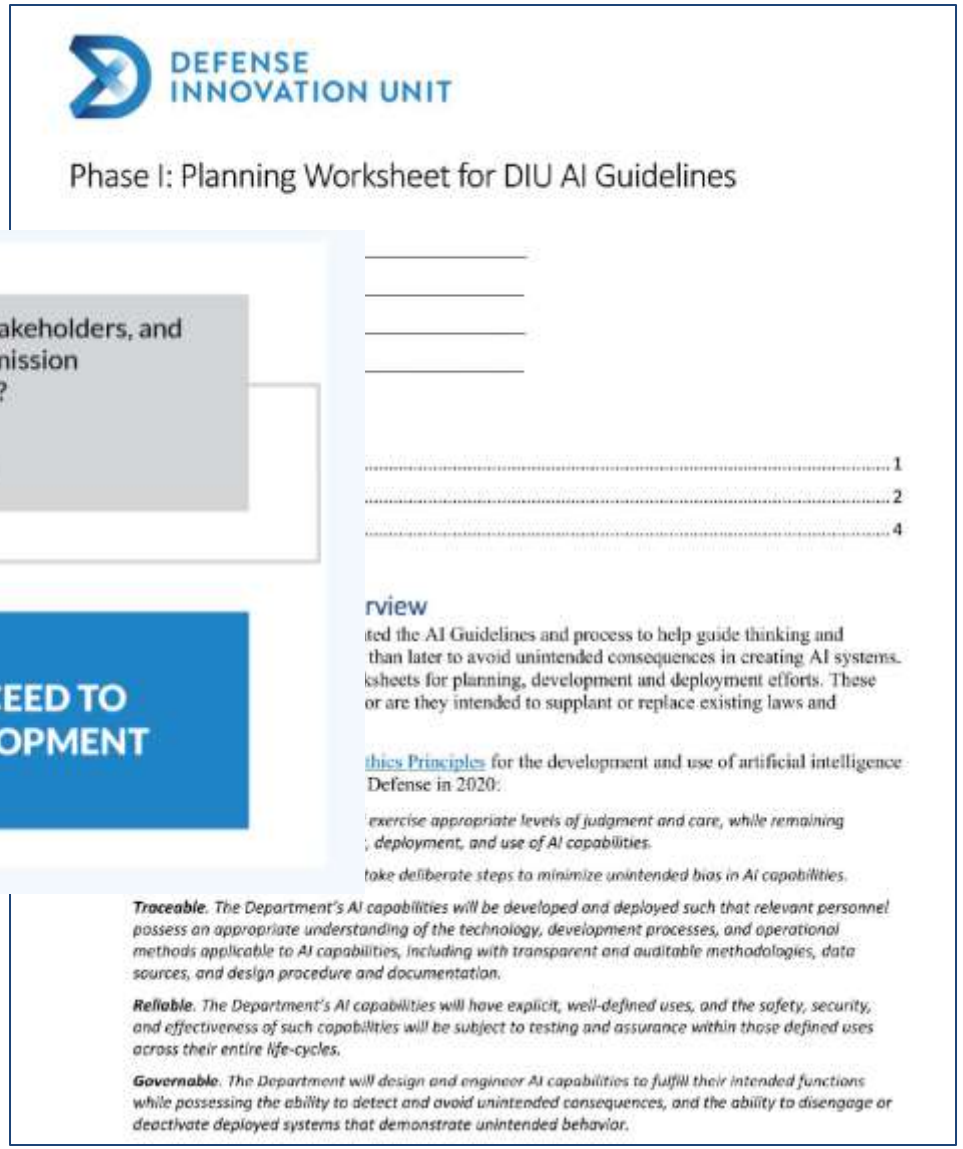
Operationalizing DoD's Ethical Principles for AI

Download DIU's Responsible AI Guidelines report and learn how to implement ethical AI principles.

Responsible AI Guidelines

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

Phase I: Planning


DEFENSE INNOVATION UNIT

Phase I: Planning Worksheet for DIU AI Guidelines

rview

ted the AI Guidelines and process to help guide thinking and than later to avoid unintended consequences in creating AI systems. ksheets for planning, development and deployment efforts. These or are they intended to supplant or replace existing laws and

thics Principles for the development and use of artificial intelligence Defense in 2020:

exercise appropriate levels of judgment and care, while remaining , deployment, and use of AI capabilities.

take deliberate steps to minimize unintended bias in AI capabilities.

Traceable. *The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation.*

Reliable. *The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.*

Governable. *The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior.*

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

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?

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

Categories of Harm (ServiceEase)

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

SmartPackage Scenario

SmartPackage - Scenario

Track online orders, shipping progress, and receipt.

Users: Consumers at home

Goals of SmartPackage:

- No worries - expected delivery is easy to track
- Alert when arrive and location via images
- Manages returns

Significant decisions

SmartPackage

- Ability to turn on and off notifications
- Ability to control camera content/capture

Responsibilities explicitly defined

Between system and human(s)

SmartPackage (System or Consumer?)

- Integrates new purchases?
- Integrates new vendors?
- Determines when to send alert?
- How many to keep available?

Respectful and Secure

SmartPackage

- Who has access to shipments and contents?
- Who has access to images of shipments and address?
- How is that information used?
- How is PII* of consumers protected?

*PII is Personally Identifiable Information (name, address, etc.)