CCSQ January 2022

Engaging with Ethics to Make Trustworthy Experiences

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

Twitter: @carologic @SEI_CMU_AI

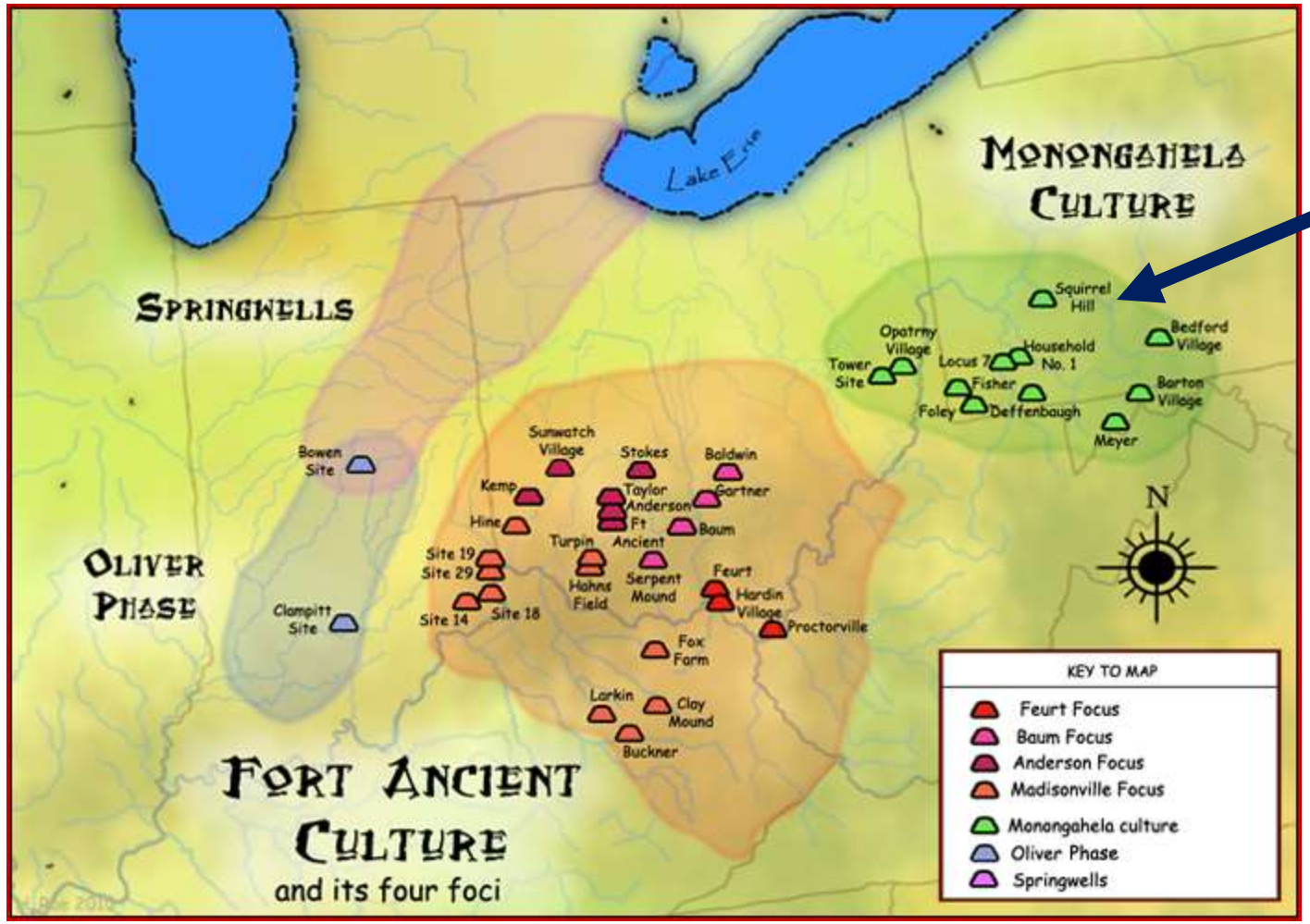
Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

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.

DM22-0094

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

AI and 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.

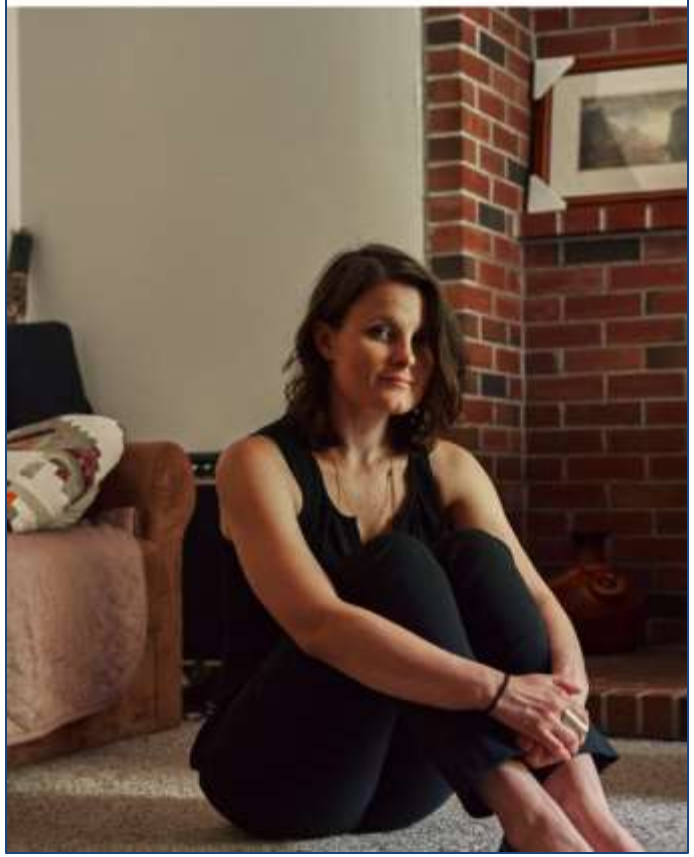
By Mark Hanrahan
December 12, 2019, 9:56 PM • 7 min read

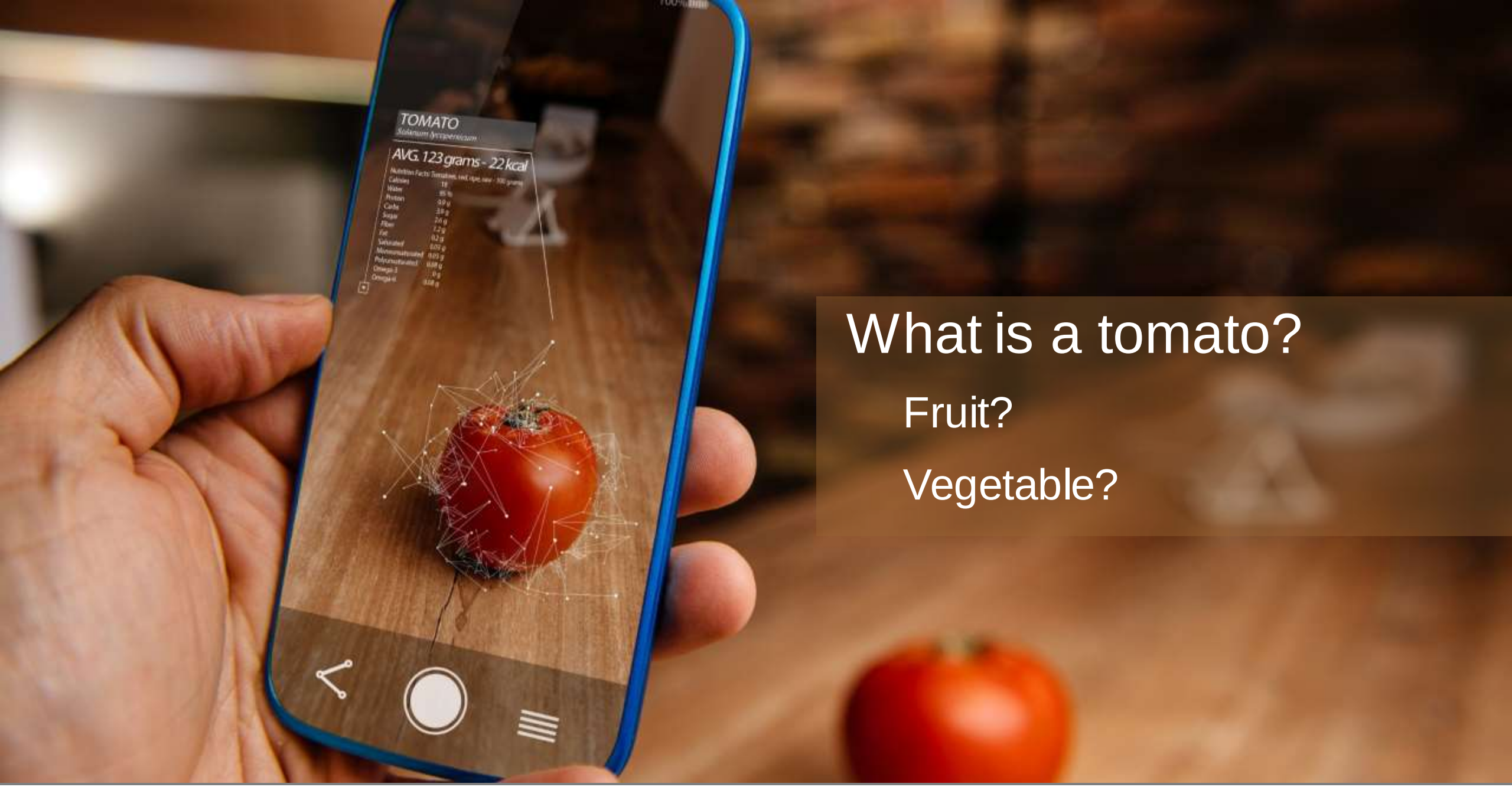
Ring camera systems being hacked

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





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

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

Responsible, Intentional Design

Just because you can, doesn't mean you should.

Ethics and UX Framework

Ethics

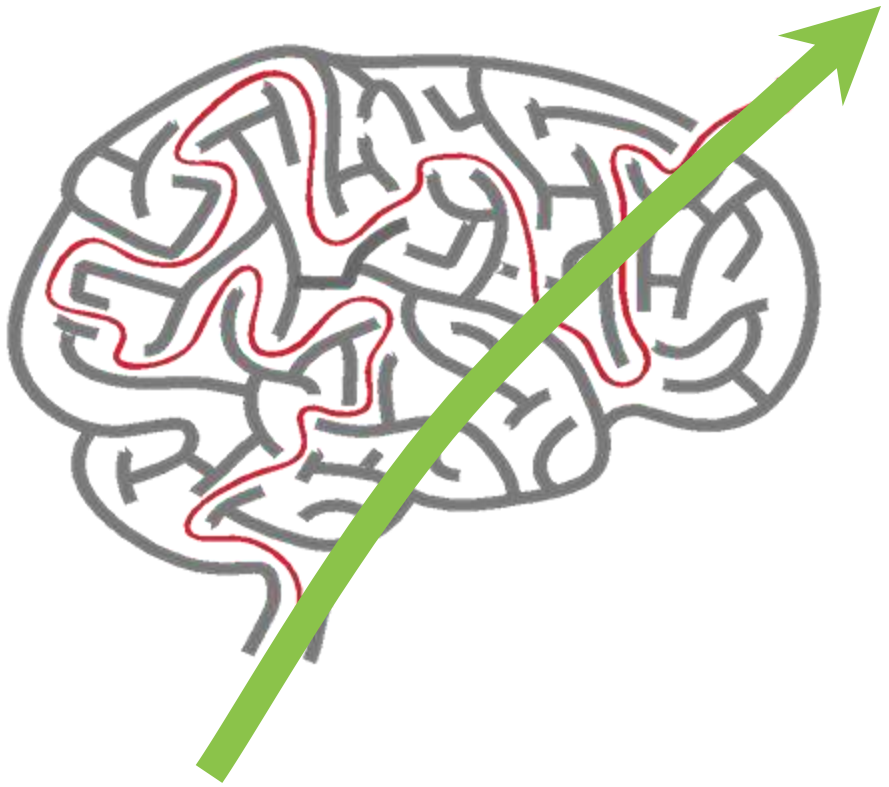
Based on well-founded standards of right and wrong

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

Talk about biases in non-threatening, productive ways

All systems have some form of bias

Complete objectivity is misleading.

Bias can have purpose and can be helpful.

The goal is to reduce unintended and/or harmful bias.

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

Adopt Technology Ethics

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

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

Early, purposeful work

In addition to the usual UX work

- Is this problem solvable with technology?
- What kind of improvements are expected?
- How will the system partner with people? Compliment?
- What are the obvious risks?
- How might these systems be misused/abused?

Leaders establish psychological safety for diverse, multi-disciplinary teams

Shared Tech Ethics

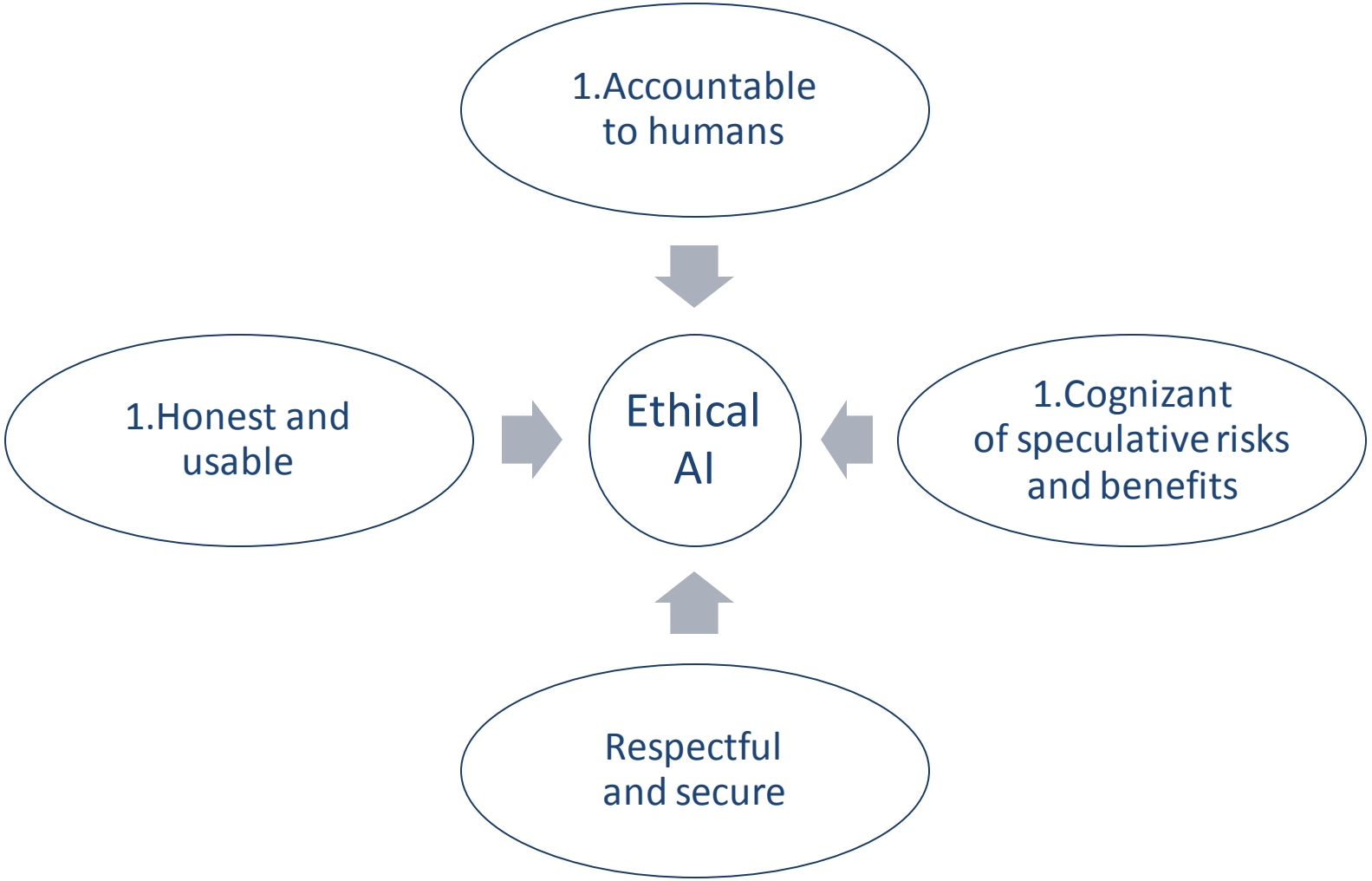
UX Framework

How do we get there?

Trustable, Ethical Systems

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

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

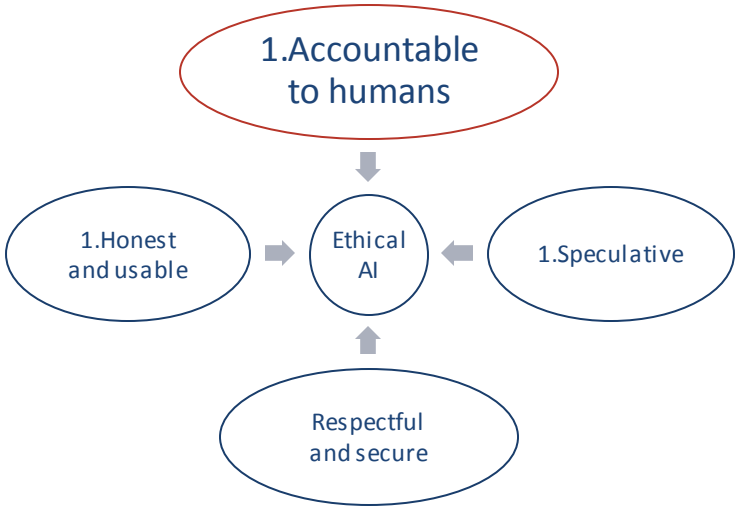
Accountable to Humans

Ensure humans have ultimate control

- Able to monitor and control risk

Human responsibility for final decisions

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



Significant decisions

Significant decisions made by the system will be

- explained
- able to be overridden
- appealable and reversible

Responsibilities explicitly defined

- Between system and human(s)

“Ensure humans can unplug the machines”

– Grady Booch

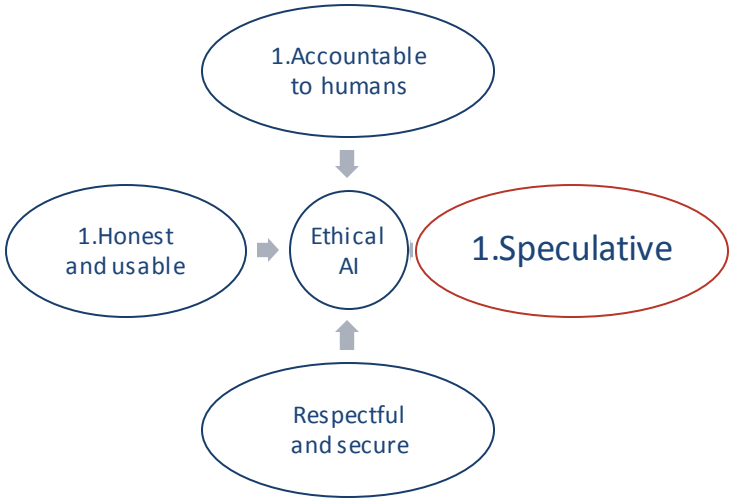
TED Talk, Grady Booch, Scientist, Philosopher, IBM’er

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

Cognizant of Speculative Risks and Benefits

Identify full range of

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



Speculative: Create communication & mitigation plans

Plan for unwanted consequences

Misuse and abuse of system

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

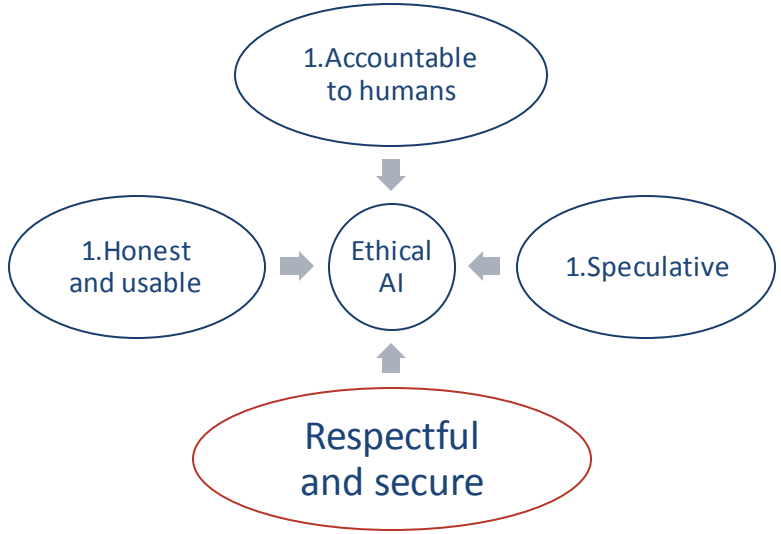
Respectful and Secure

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

Respect privacy and data rights

Make system robust, valid and reliable

Provide understandable security



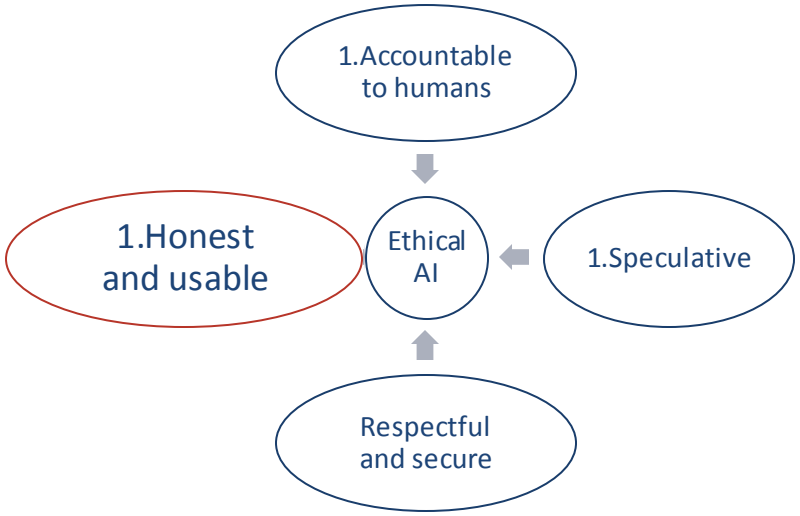
Honest and Usable

Value transparency with the goal of engendering trust

Explicitly state identity as an AI system

Show awareness of known and desirable bias

Acknowledge and overcommunicate issues



Conversations for Understanding

Conversations for Understanding

Difficult Topics

- What do we value?
- Who could be hurt?
- What lines won't our system 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

Principles and 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

U.S. DEPT OF DEFENSE

Montréal Declaration Responsible AI_

An initiative of Université de Montréal

Checklist

Pair Checklist with Technical Ethics

- Bridges gap between "do no harm" and reality

Drives conversations

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 consent 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

New uncomfortable work

“*Be uncomfortable*”

- Laura Kalbag

Ethical design is not superficial.

SmartPackage

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

SmartPackage (System or Consumer?)

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

Respectful and Secure

SmartPackage

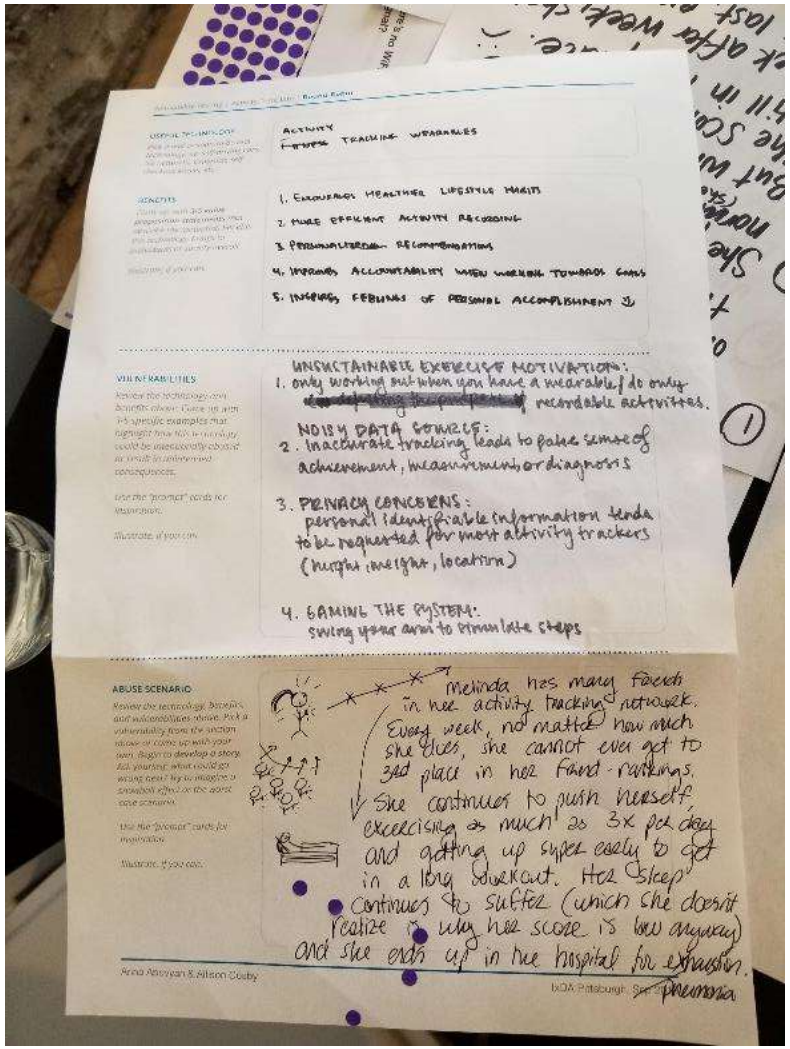
- Who has access to shipments and contents?
- Who has access to images of shipments and address?
- How is that information used?
- How is PII* of consumers protected?

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

Curiosity Inspiring Activities

Abusability Activity

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



UX research methods for ethics

- 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

Abusability Testing

Feature added to enable SmartPackage to report delivery issues

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

“Black Mirror” episode

- SmartPackage was designed for individual use.
- Households with multiple people receiving packages, find that SmartPackage starts to refuse deliveries meant for other household members
- SmartPackage reports them as errors to the shipping carriers, creating further frustration and grief
- And then…

Overview

Pick a newish technology:

Self-checkout kiosks, human augmentation (implants), home video monitoring, drone package delivery, smart watches, 5G networks, social media, self-driving cars, etc. etc.

Brainstorm:

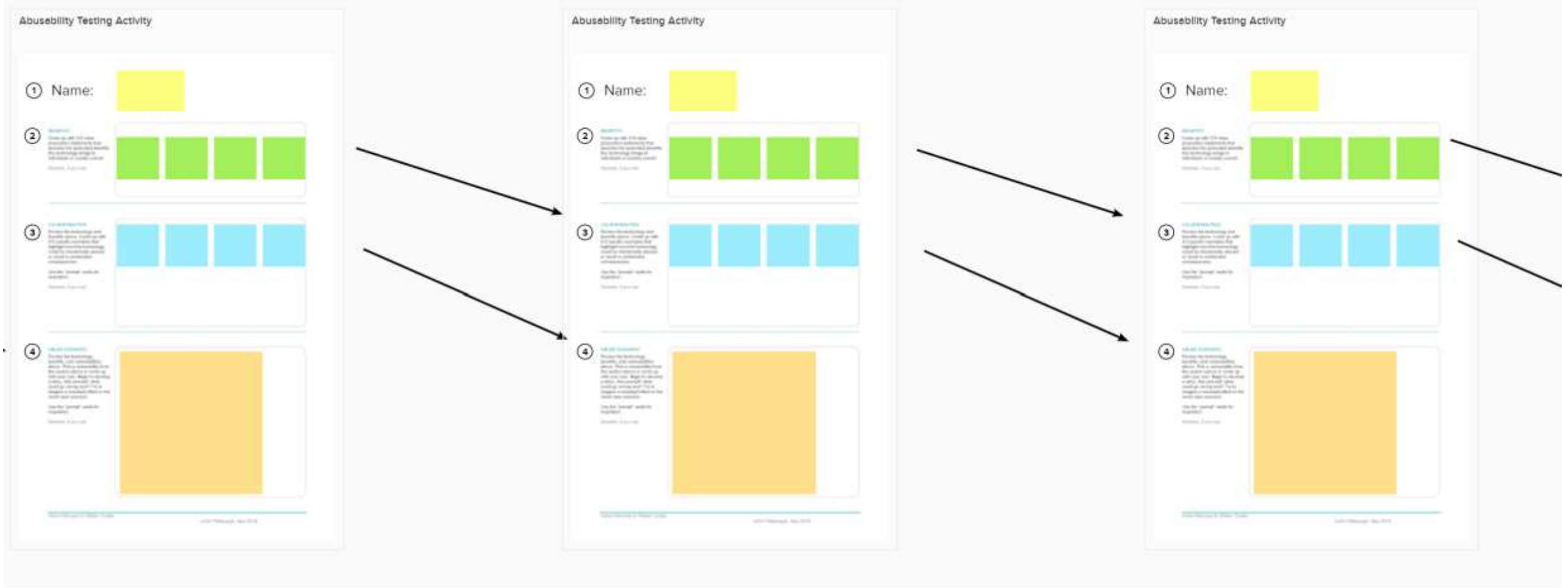
- Benefits
- Vulnerabilities
- Abuse Scenario
- "Black Mirror" Episode Scenario (Casey Fiesler)

  (inspired by British dystopian sci-fi tv series of same name)

Prompt Statements

- What happens if the electricity goes out?
- What happens if devices can charge themselves?
- What if there's no WiFi or cellular signal?
- You've been hacked.
- How is status communicated?
- Is there a back door?
- What error conditions need to be considered?

Abusability Testing



DEFENSE INNOVATION UNIT

ARTIFICIAL INTELLIGENCE PORTFOLIO

Responsible AI Guidelines

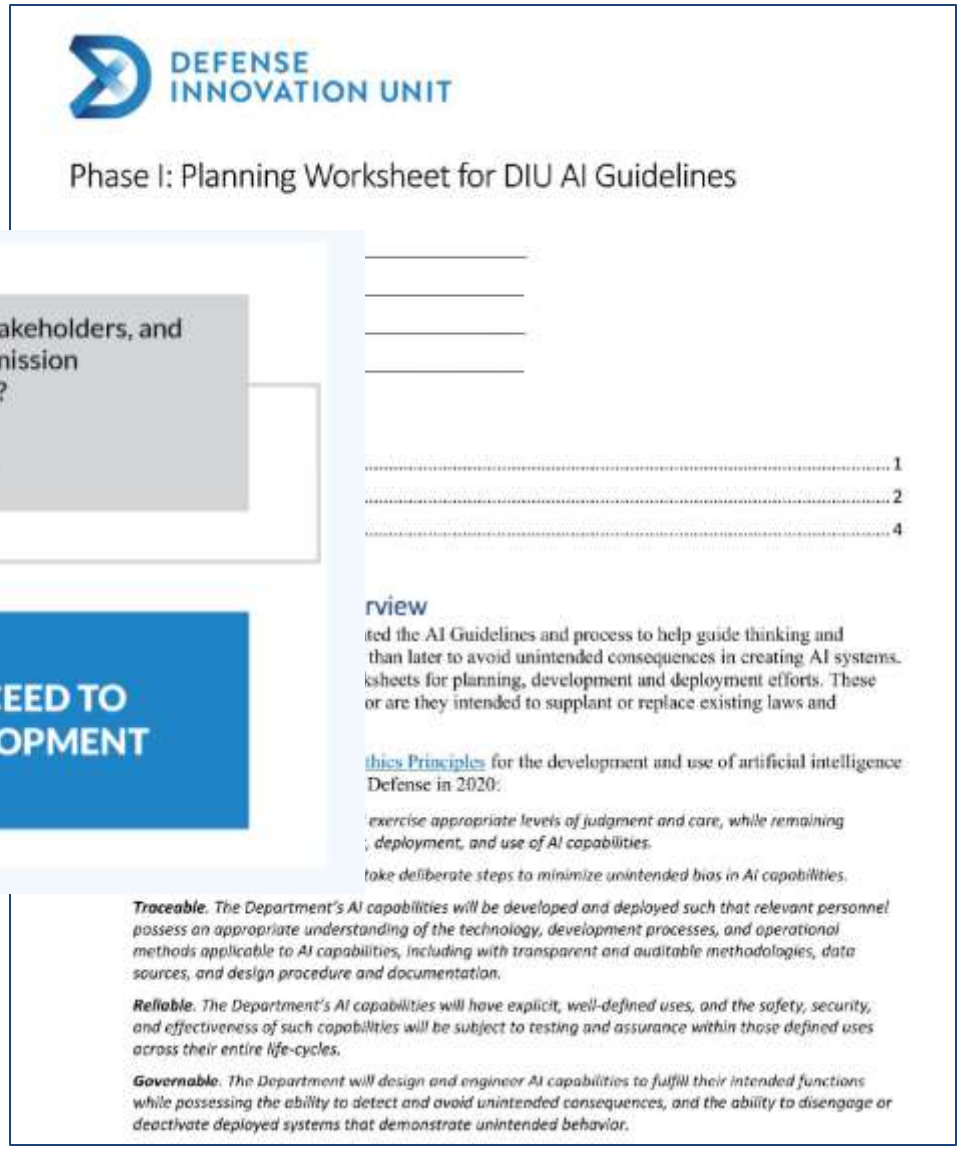
Operationalizing DoD's Ethical Principles for AI

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

Taking Next Steps

Tools to take you forward

- Set of ethical principles
- UX Framework and checklist to encourage conversations for understanding
- Method: Abusability Testing

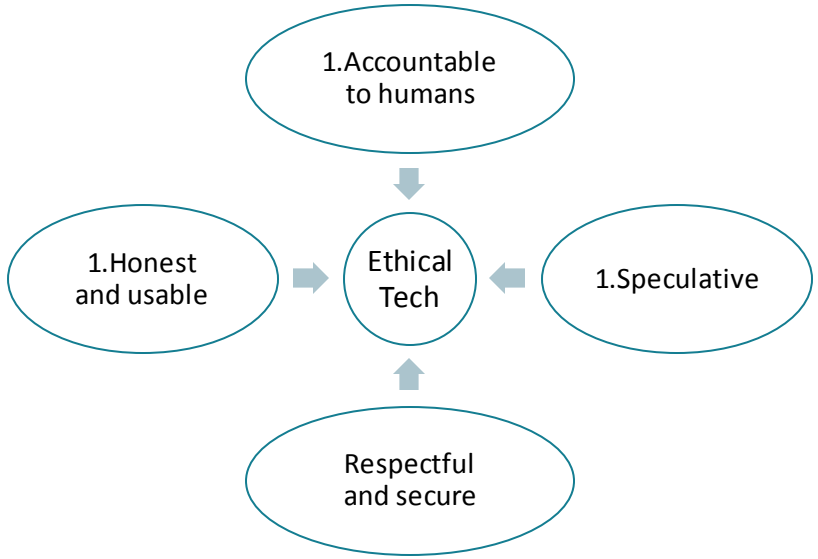
We aren’t perfect, tech won’t be perfect

Empower diverse teams, inclusive environments

Adopt technical ethics

Encourage deep conversations (Checklist)

Activate curiosity; be speculative; imaginative



Evangelize for human values

Ethical. Transparent. Fair.

Carol J. Smith

Twitter: @carologic

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

CMU Software Engineering Institute,
AI Division

Twitter: @SEI_CMU_AI

Discussion

Topics

Why does UX have to manage ethics?

Why activate curiosity?

What scenarios do you have?

- At work?
- In your personal life?
- In the news?
- Where else would you want to apply this?

Initial organizational questions

- What value do we see in defining a set of ethical guidance for our organization?
- Should we build on existing codes, guidance and best practices?
- Would different departments/products need different guidance?
- How might we reflect our culture/what we value in a code of ethics?
- Who do we need to protect?
- How will we describe the “lines” our work won’t cross?
- What mitigation strategies do we need?
- What about social consequences such as job displacement?
- How might we track progress?

Resources

Resources

AI Now Institute - https://ainowinstitute.org/
Medium article: After a Year of Tech Scandals, Our 10 Recommendations for AI: Let's begin with better regulation, protecting workers, and applying "truth in advertising" rules to AI https://medium.com/@AINowInstitute/after-a-year-of-tech-scandals-our-10-recommendations-for-ai-95b3b2c5e5

ACM Code of Ethics and Professional Conduct - The Official Site of the Association for Computing Machinery's Committee on Professional Ethics: https://ethics.acm.org/ Adopted by Association for Computing Machinery's Council 6/22/18

Google: our principles - https://blog.google/technology/ai/ai-principles/

IDEO Medium article: Data, Ethics, and AI: Practical activities for data scientists and other designers. By Michael Chapman, Ovetta Sampson, Jess Freaner, Mike Stringer, Justin Massa, and Jane Fulton Suri https://medium.com/ideo-stories/data-ethics-and-ai-276723a1a2fc

IBM Medium Article: Everyday Ethics for Artificial Intelligence: a Practical Guide for Designers and Developers by Adam Cutler, IBM Distinguished Designer, Artificial Intelligence Design; Milena Pribić, IBM Designer, Artificial Intelligence Design; and Lawrence Humphrey, IBM Designer, Artificial Intelligence Design: https://medium.com/design-ibm/everyday-ethics-for-artificial-intelligence-75e173a9d8e8
Everyday Ethics for Artificial Intelligence: A Practical Guide for Designers and Developers, IBM, 2018
https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

Microsoft AI Principles - https://www.microsoft.com/en-us/ai/our-approach-to-ai

Additional Resources

Algorithm Tips - Resources: http://algorithmtips.org/resources/

Ethics and Data Science by DJ Patil, Hilary Mason, Mike Loukides. Publisher: O'Reilly Media, Inc.. Release Date: July 2018. ISBN: 9781492043898. https://www.oreilly.com/library/view/ethics-and-data/9781492043898/

Ethical OS: https://ethicalos.org/

PAPA (Privacy, Accuracy, Property, Accessibility) - Ethical Issues in IS by Richard Mason. https://www.gdrc.org/info-design/4-ethics.html

Tech Ethics Curricula: A Collection of Syllabi. Casey Fiesler (Faculty in Information Science at CU Boulder). Jul 5, 2018. https://medium.com/@cfiesler/tech-ethics-curricula-a-collection-of-syllabi-3eedfb76be18

Toward ethical, transparent and fair AI/ML: a critical reading list. By Eirini Malliaraki http://emalliaraki.com/ Feb 19, 2018. https://medium.com/@eirinimalliaraki/toward-ethical-transparent-and-fair-ai-ml-a-critical-reading-list-d950e70a70ea

UXPA Code of Professional Conduct: https://uxpa.org/resources/uxpa-code-professional-conduct

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/