World Usability Day 2020
Cleveland, Ohio

Implementing Ethics in Emerging Technologies

Carol J. Smith
Sr. Research Scientist - Human-Machine Interaction

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.

DM20-1051

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.

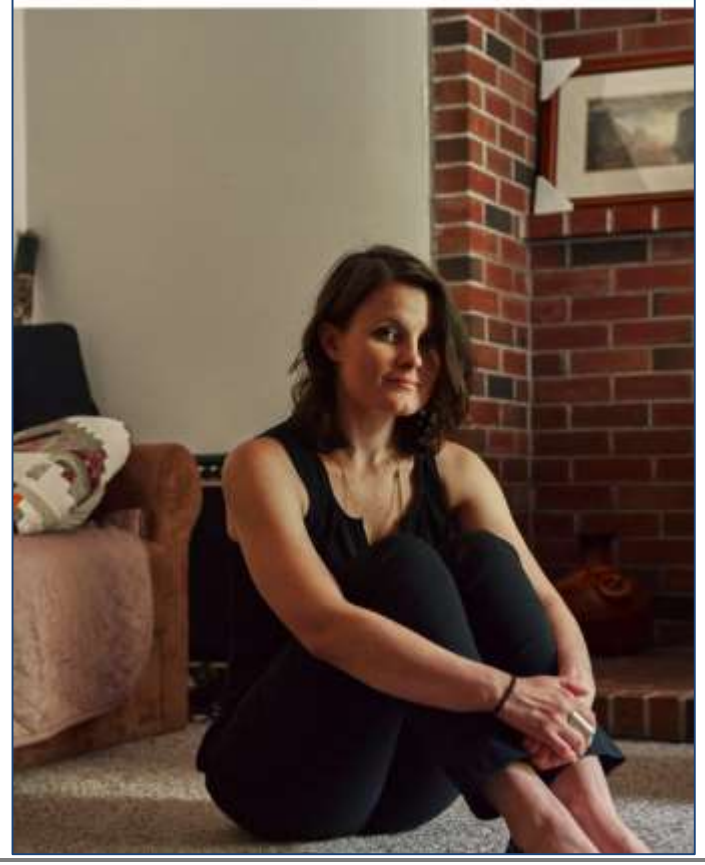
By Mark Hanrahan
December 12, 2019, 9:56 PM • 7 min read

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



FAST COMPANY

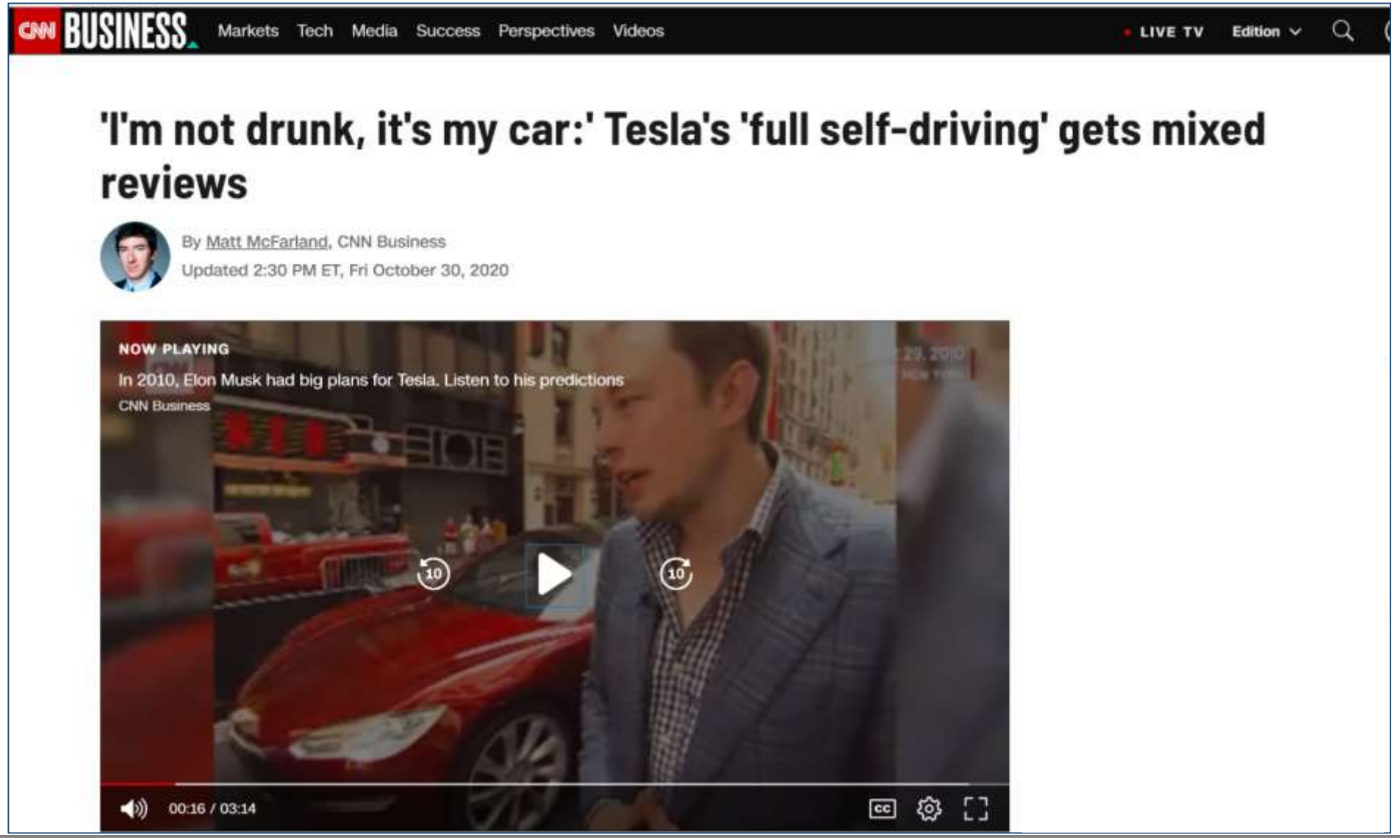
07-07-17

Nest Founder: "I Wake Up In Cold Sweats Thinking, What Did We Bring To The World?"

Tony Fadell, one of the minds behind the iPod and the iPhone, mulls design's unintended consequences.

[Photos: Constantin Renner/EyeEm/Getty Images, davide ragusa/Unsplash]

Just ADAS - #Autonowashing



Responsible, Intentional Design

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

Early, purposeful work

In addition to the usual UX work

- What kind of improvements are expected with the emerging technology?
- 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

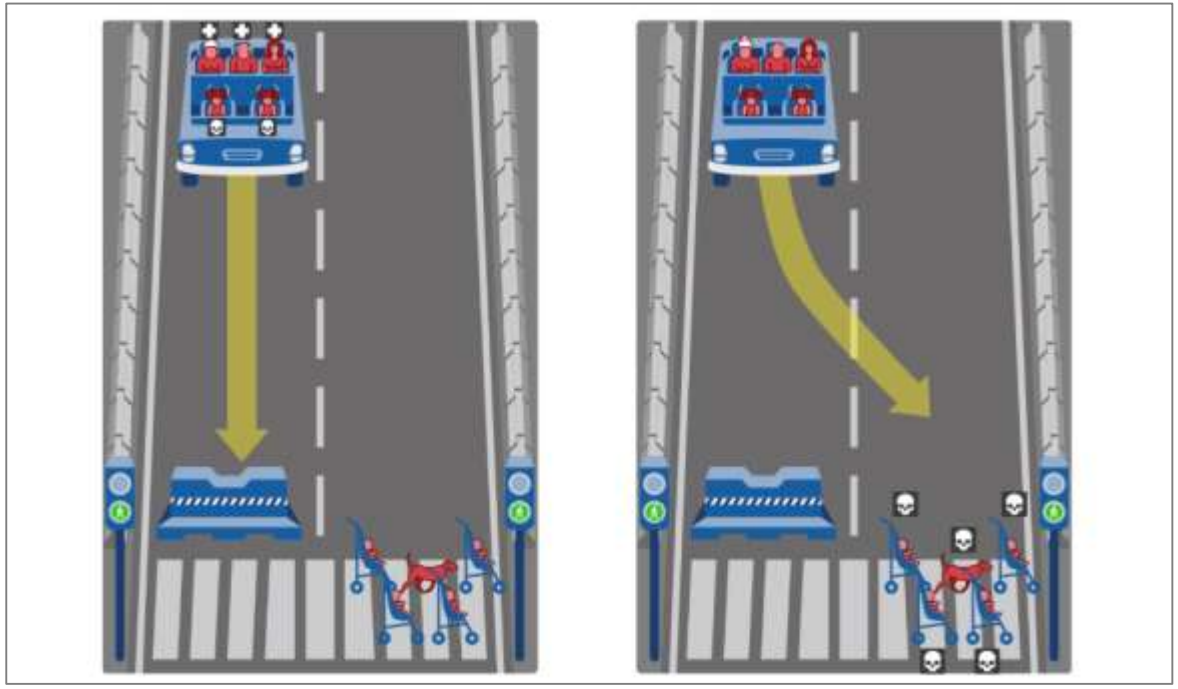
Standard of expected behavior that guides the correct course of action

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/

Not Trolley Problems (hypotheticals in-the-moment)

MIT Moral Machine Project



Does the Trolley Problem Have a Problem?

What if your answer to an absurd hypothetical question had no bearing on how you behaved in real life?

By DANIEL ENGBER JUNE 18, 2018 • 5:56 AM

Lisa Larson-Walker

Images: 1) MIT Moral Machine Project: http://moralmachine.mit.edu/
2) Does the Trolley Problem Have a Problem? https://slate.com/technology/2018/06/psychologys-trolley-problem-might-have-a-problem.html

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

Biased due to…

Social class

Resource availability

Education

Race, gender, sexuality

Culture, theology, tradition

More…

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.

Teams - Diverse, talented and multi-disciplinary

Bringing their varied skills sets, problem framing approaches, and knowledge together.

- Gender, race, culture
- Education (school, program, etc.)
- Experiences
- Thinking process, skill set
- Age, disability and health status, and more…

Representatively diverse leadership for retention

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

Not lowering bar

——— extending it

Great Minds Think Different

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 — entrenched ways of thinking that can otherwise blind them to key information and even lead them to make errors in decision-making processes."

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

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

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

ETHICAL OS

Diverse, inclusive leaders

Diverse, Multi-Disciplinary Teams

Shared Tech Ethics

UX Framework

Implementing Ethics

Activate curiosity

UX research methods to activate curiosity:

- Abusability Testing
- “Black Mirror” Episodes (inspired by British dystopian sci-fi tv series of same name)
- Flip it to test it
- Implicit Association Test from Harvard

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

How do we get there?

Trustable, Ethical Systems

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

Conversations for Understanding

UX Framework guides teams

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

New uncomfortable work

“*Be uncomfortable*”

- Laura Kalbag

Ethical design is not superficial.

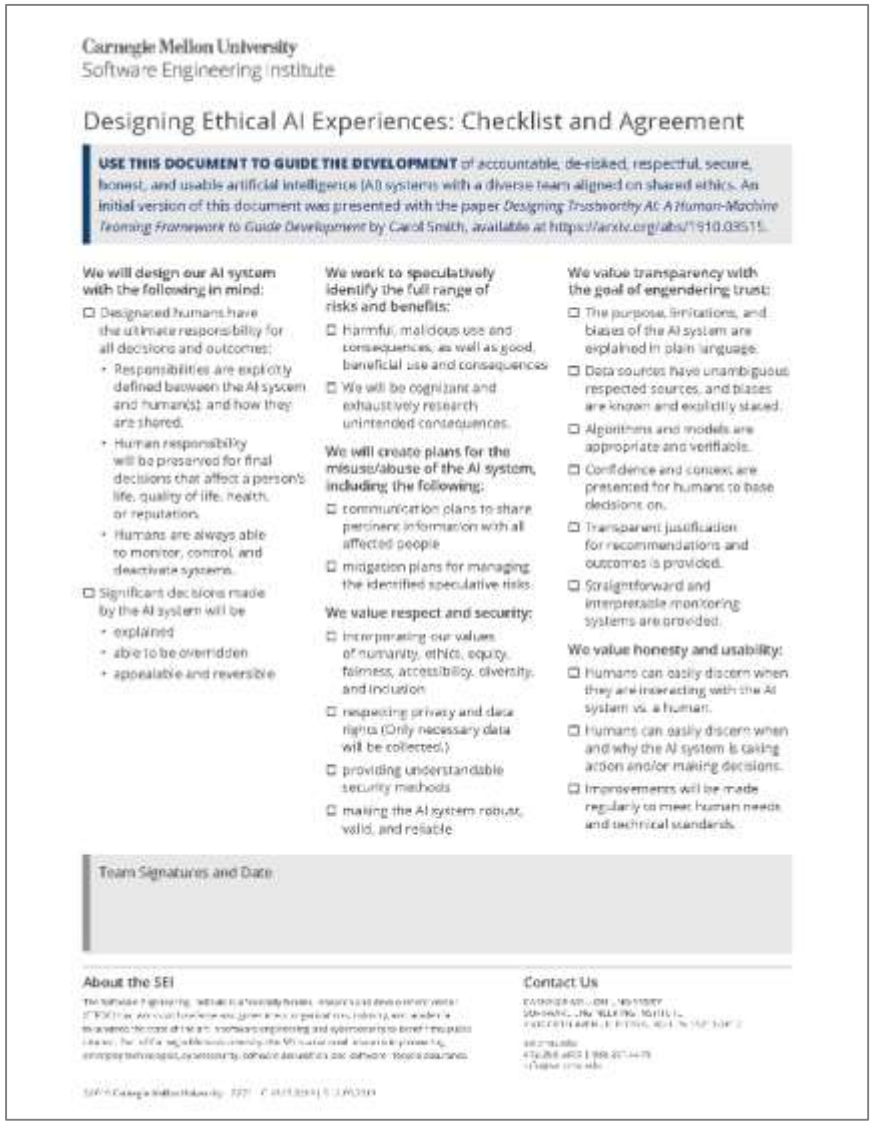
Prompt conversations

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 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.

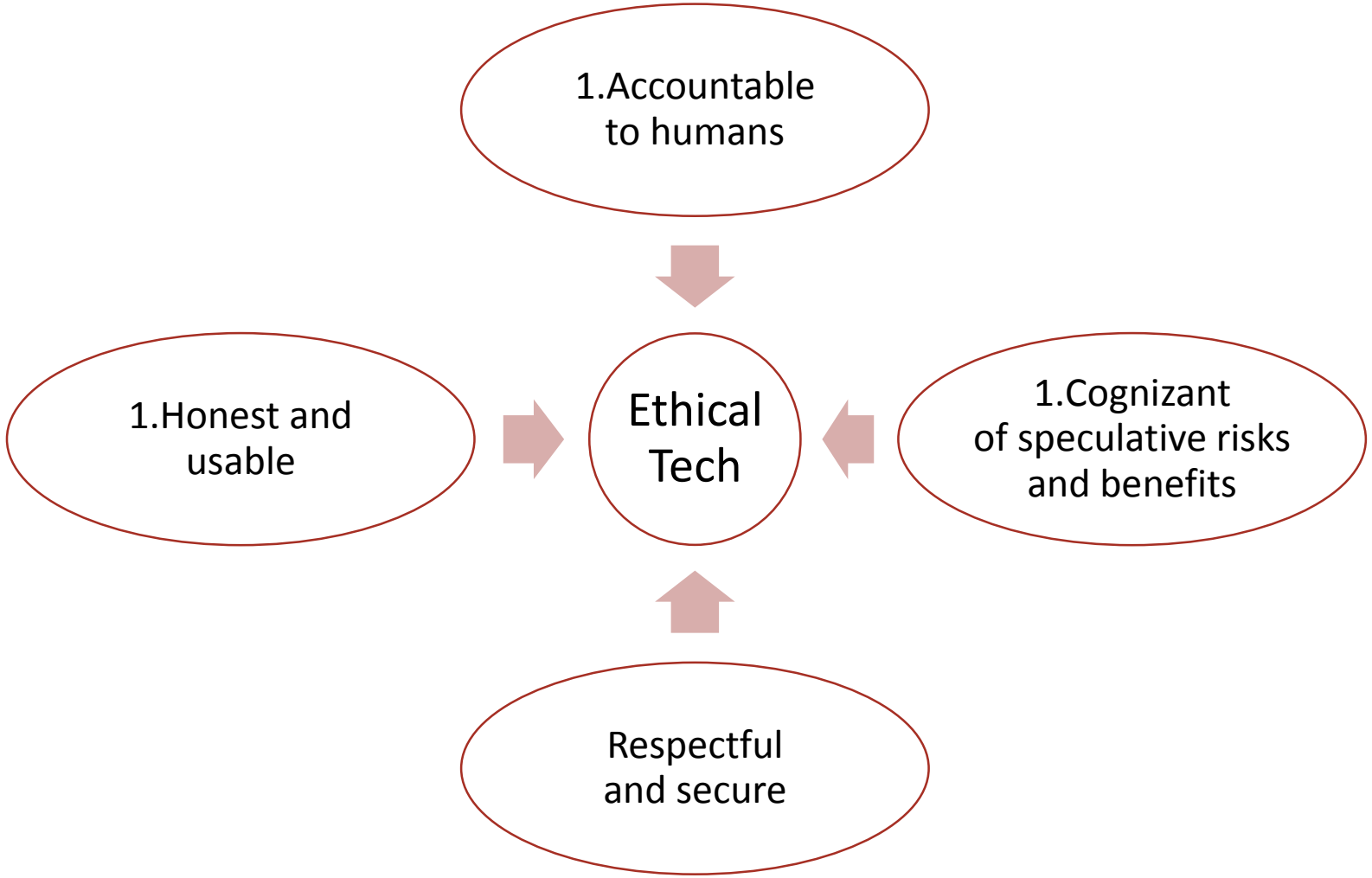
Team Signatures and Date

About the SEI

Contact Us

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

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

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

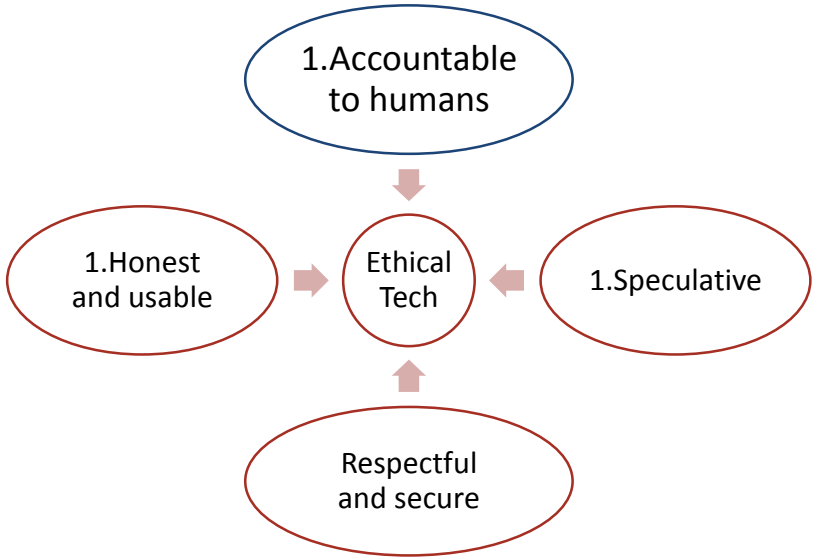
Accountable to Humans

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 will be

- explained
- able to be overridden
- appealable and reversible

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?

Abusability Testing

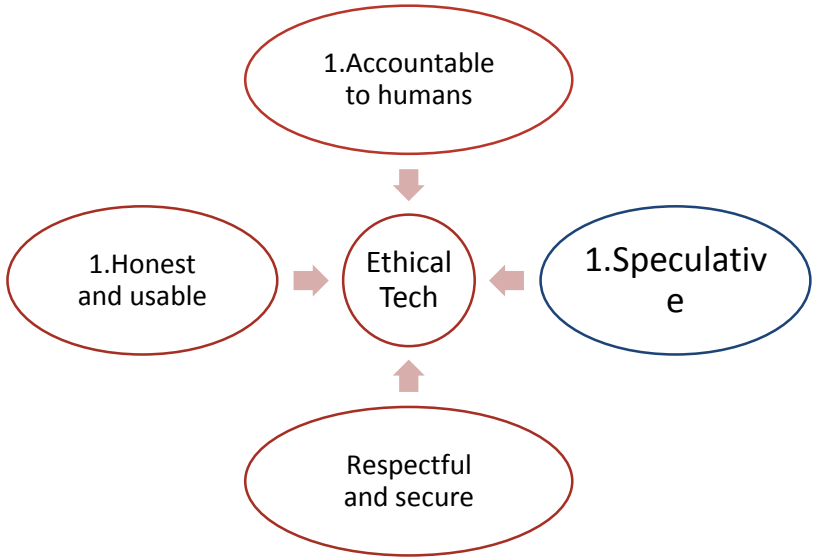
Feature added to enable SmartPackage to report delivery issues

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

Cognizant of Speculative Risks and Benefits

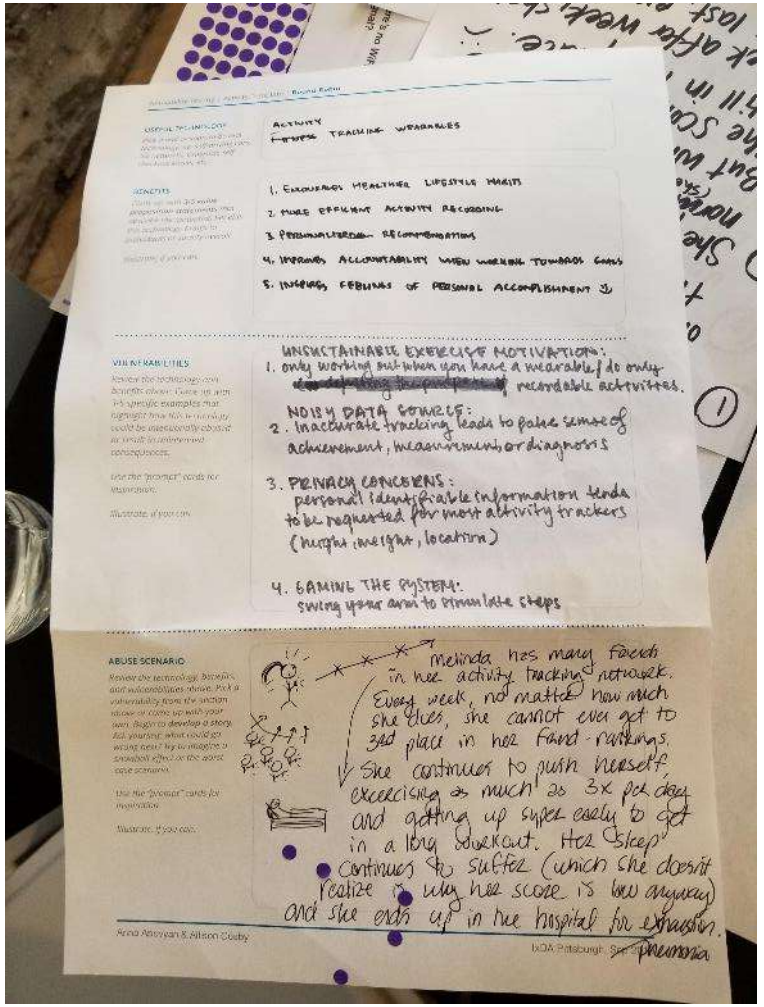
Identify full range of

- Harmful, malicious use, as well as good, beneficial use
- Blind spots and 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

- 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

Speculative: Create communication & mitigation plans

Plan for unwanted consequences

Misuse and abuse of system

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

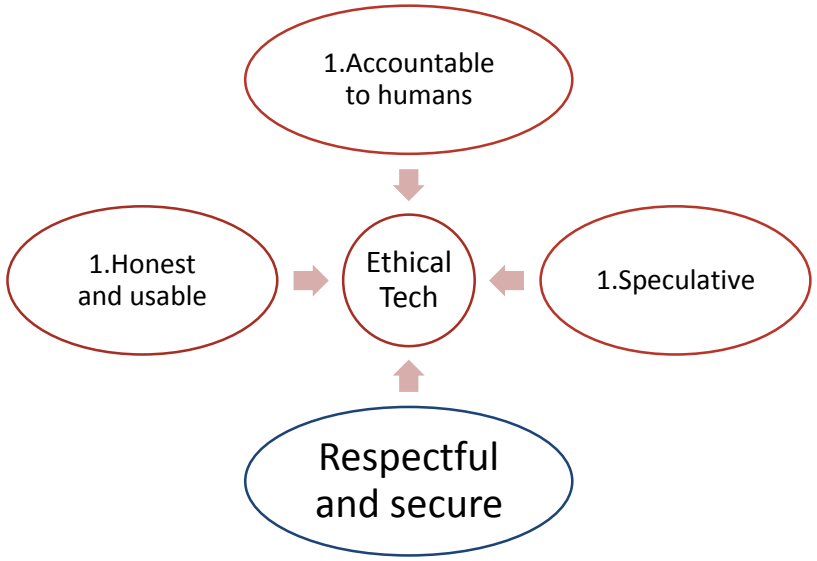
Respectful and Secure

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

Respect privacy and data rights

Make system robust, valid and reliable

Provide understandable security



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.)

Honest and Usable

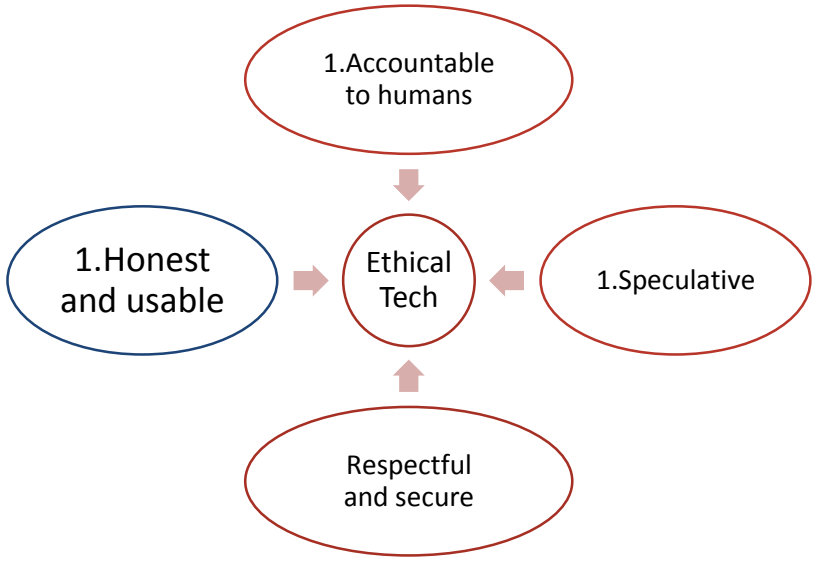
Value transparency with the goal of engendering trust

Smart speakers explicitly state identity as an AI system

Remove unwanted bias in data

Show awareness of known and desirable bias

Acknowledge issues



Reward team members for finding ethics bugs

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

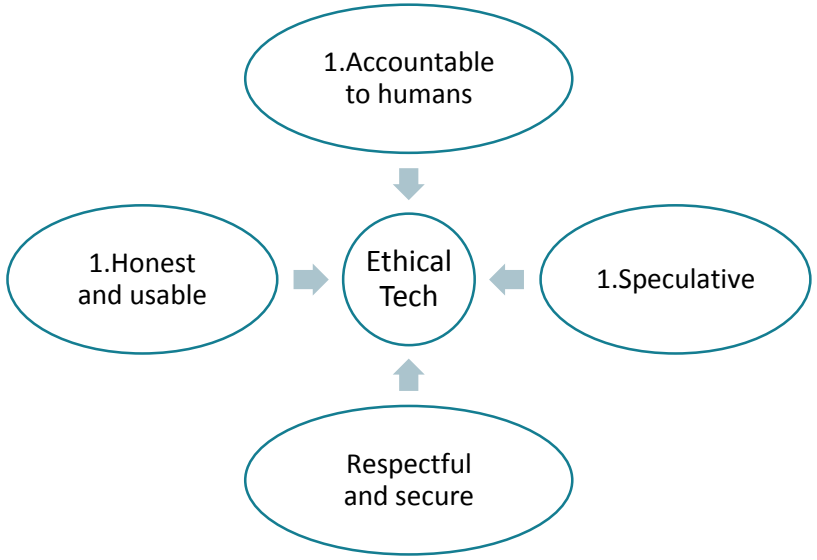
We aren’t perfect, tech won’t be perfect

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's Software Engineering Institute,
Emerging Technology Center

Twitter: @sei_etc