Workshop

Ethical Design Matters

Carol J. Smith & Lisa D. Dance
June 2021

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.

DM21-0604

Acknowledgement: The Land We Speak On

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

Now known
as Pittsburgh, PA, USA.

The land of the Powhatan
Confederacy

Now known
as Richmond, VA, USA.

Welcome!

15 min - Opening

20 min - Intro to ethics and technology

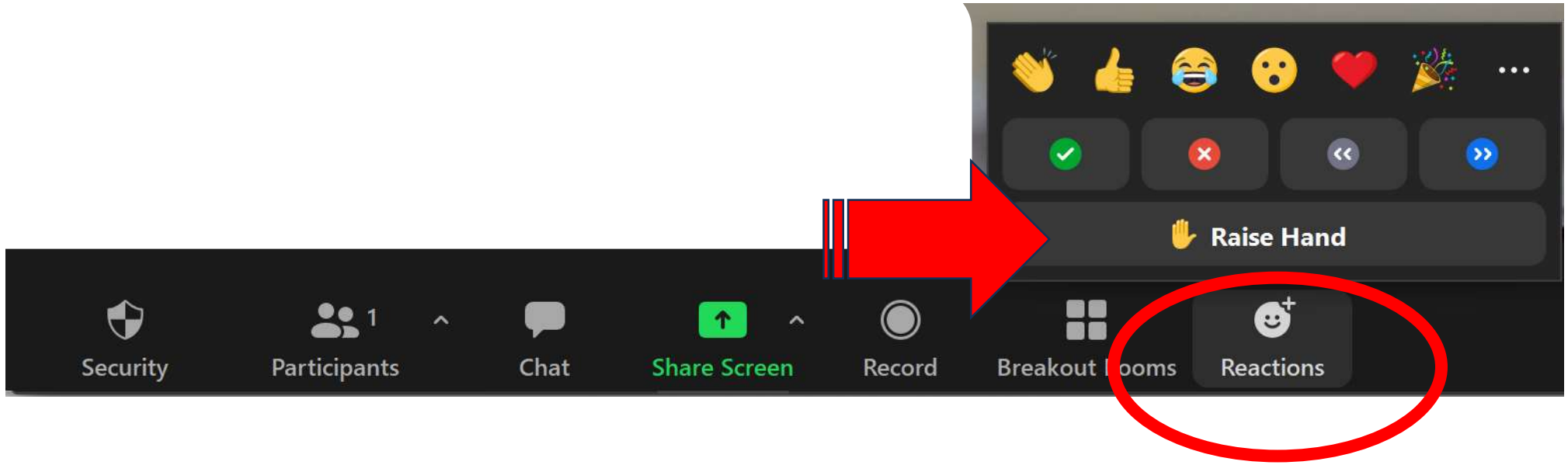
60 min - Activity

25 min - Discussion / Q&A

All times are approximate

Be patient and kind - technology will likely fail

Using Zoom



Why does ethical design matter?

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.

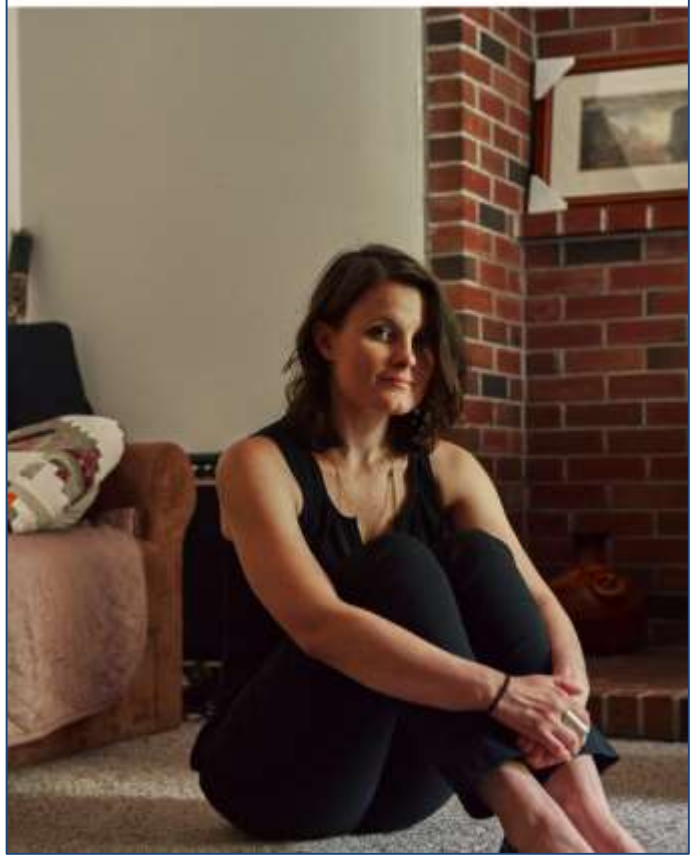
By Mark Hanrahan
December 12, 2019, 9:56 PM • 7 min read

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



Assuming That Math Reduces Bias

Vetting applicant resumes

REUTERS

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / 10 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Embedding Existing Bad Behaviors

Determining interest rates for mortgage lenders

Dehumanizing thru Math

Predicting future criminal activity

Ignoring Systemic Inequalities

Determining follow up care

PBS NEWS HOUR WEEKEND

Widely used algorithm for follow-up care in hospitals is racially biased, study finds

Ignoring Abusability

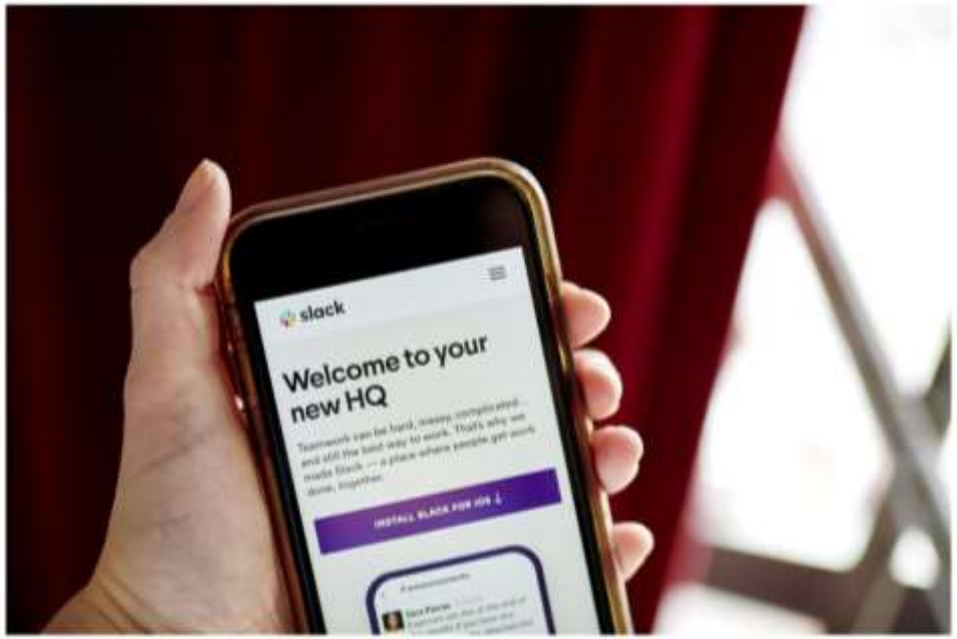
Opting in everyone to messaging

THE WALL STREET JOURNAL.

TECH

Slack Backtracks on New Way to Message Over Harassment Concern

Workplace chat pioneer says it made a mistake in how it rolled out direct-messaging function



Slack use traditionally has been limited to chat within an organization.

PHOTO: GABBY JONES/BLOOMBERG NEWS

Perpetuating Bias

Determining gender by name

Service that uses AI to identify gender based on names looks incredibly biased

Meghan Smith is a woman, but Dr. Meghan Smith is a man, says Genderify

By James Vincent | Jul 29, 2020, 6:44am EDT

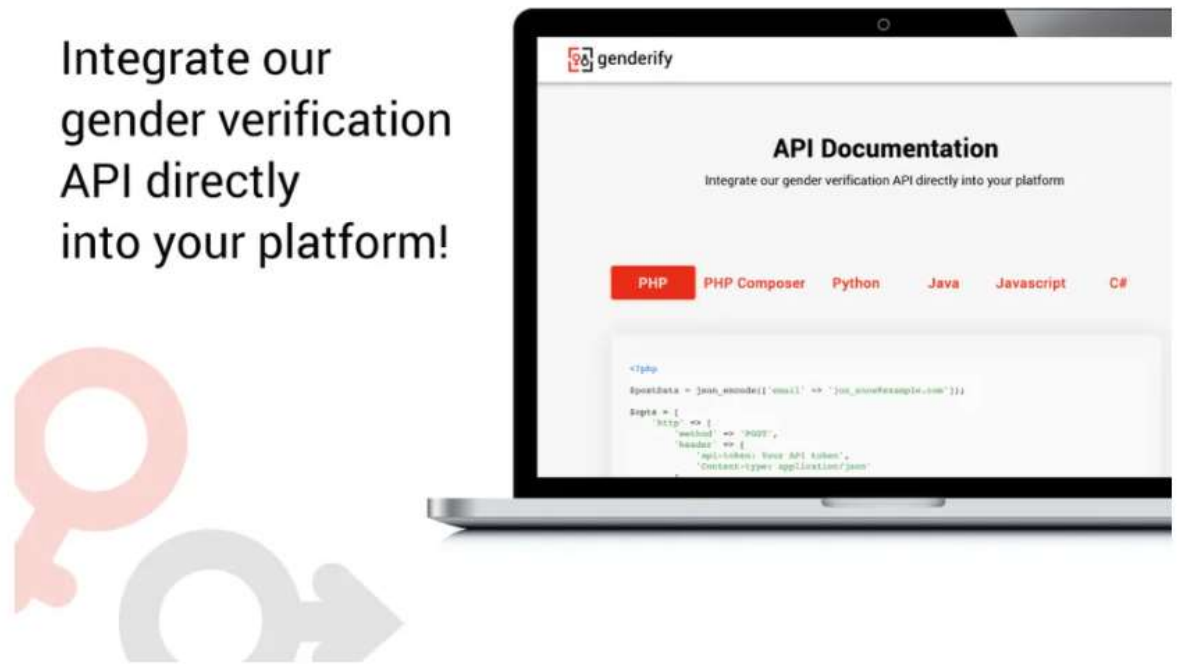


Image: Genderify

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

Responsible, Intentional Design

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

Early, purposeful work

What is the challenge being faced?

For whom? What are their needs?

What kind of improvements are expected?

What might a machine do better or faster?

What is not going to be improved (out of scope)?

Humans Make Technology

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.

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

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

Diverse, inclusive leaders

Diverse, Multi-Disciplinary Teams

Shared Tech Ethics

Frameworks & Checklists

Conversations for Understanding

Guide teams through 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.

Frameworks & checklists to prompt conversations

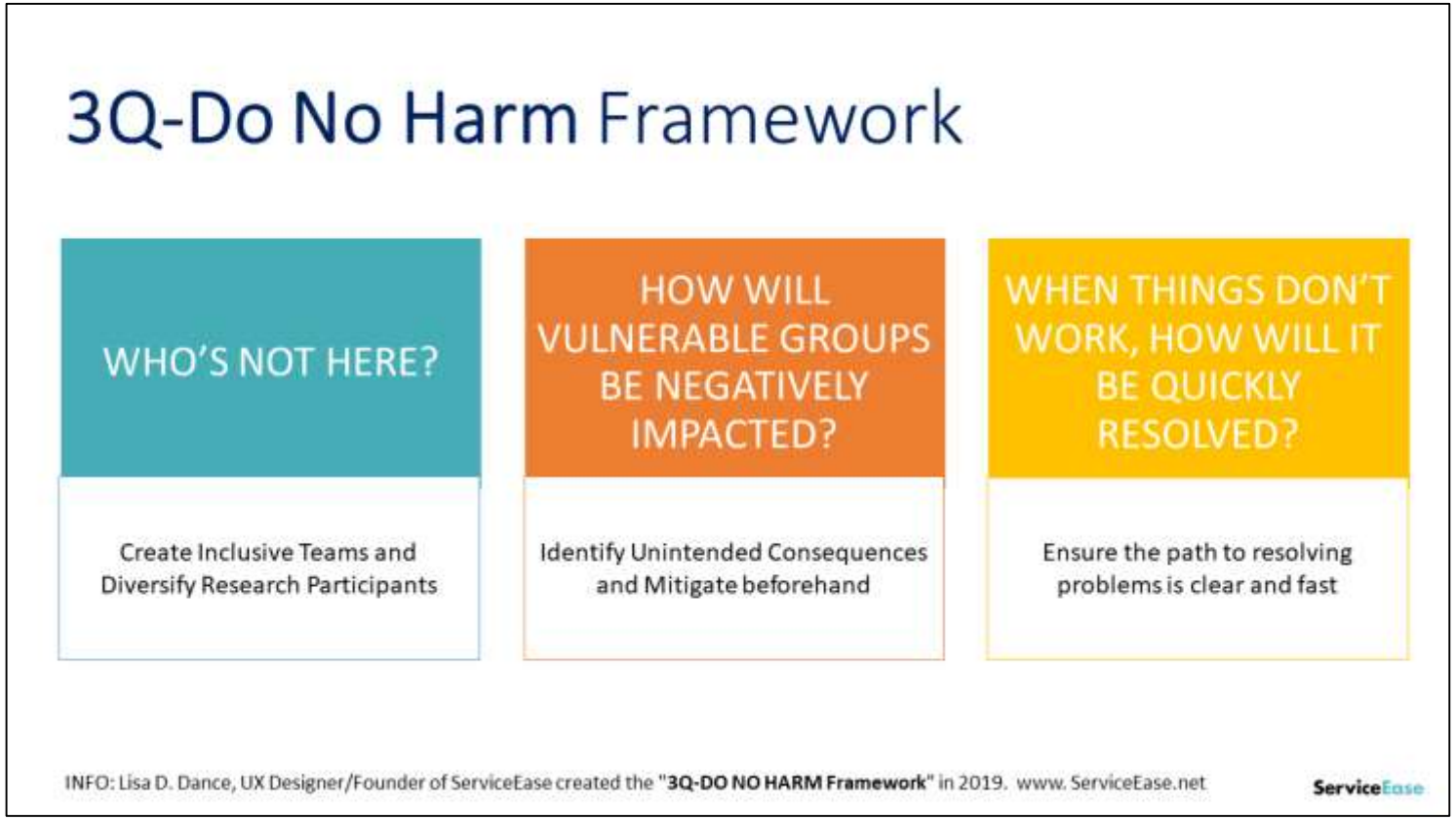
Pair with technical ethics

- Bridge gaps between “do no harm” and reality

Reduce risk and unwanted bias

Support inspection and mitigation planning

3Q-Do No Harm Framework



Framework addresses:

1) Lack of diverse perspectives
2) No examination of what harm or misuse could occur
3) Lack of consideration of what could go wrong within the context of customers' life
4) Lack of responsibility for the problems created
5) Poor implementation and poor problem resolution

Designing Ethical AI Experiences: Checklist and Agreement

UX Framework

1) Accountable to humans

2) Cognizant of speculative risks and benefits

3) Respectful and secure

4) Honest and usable

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

Categories of Harm

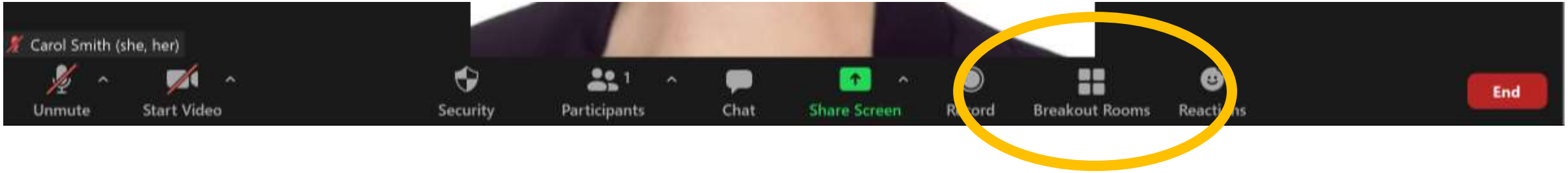
Use these categories of harm to evaluate how a product, service, or technology could cause harm.

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

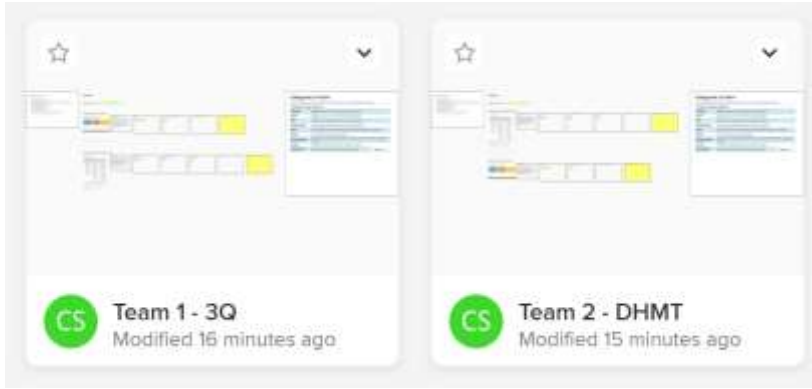
Activity in Mural

Breakout Rooms and Mural

1. View Zoom Breakout Room Assignment (Team 1, 2, 3)



2. Open Mural
3. Look for Mural square that matches Zoom Breakout Room (Team 1…)
   1. Choose to be a “guest” in Mural
   2. New to Mural? Review hints in Mural



Activity Instructions

1. Team member introductions
2. Read scenario
3. As individuals, write sticky notes to answer prompts from the specified checklist or framework
   - Either the "3Q-Do No Harm Framework“ or the "Designing for Ethical AI Checklist“
   - Use "Categories of Harm" to generate ideas
4. Discuss as a group
   - Questions, themes, Aha moments, etc.
5. Identify top 3 priorities to address
   - Severity, frequency, risk, etc.
6. Take a 5-minute break as a team
7. Repeat steps 3, 4 & 5 using the second option

* For the discussion/share out decide on a spokesperson for the team.

Activity

Team members:

3Q-Do No Harm Framework

Take a break as a team between frameworks/checklists

Designing for Ethical AI Checklist

Sharing Time

Evangelize for human values

Ethical. Transparent. Fair.

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

Empower diverse teams, inclusive environments

Adopt technical ethics

Encourage deep conversations (Checklist)

Activate curiosity; be speculative; imaginative

Carol J. Smith

Twitter: @carologic

LinkedIn: /in/caroljsmith/

&

Lisa D. Dance

Twitter: @ServiceEase

LinkedIn: /in/ldance/

Website: www.ServiceEase.net

Resources

Activate curiosity

UX research methods 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

deon: An ethics checklist for data scientists

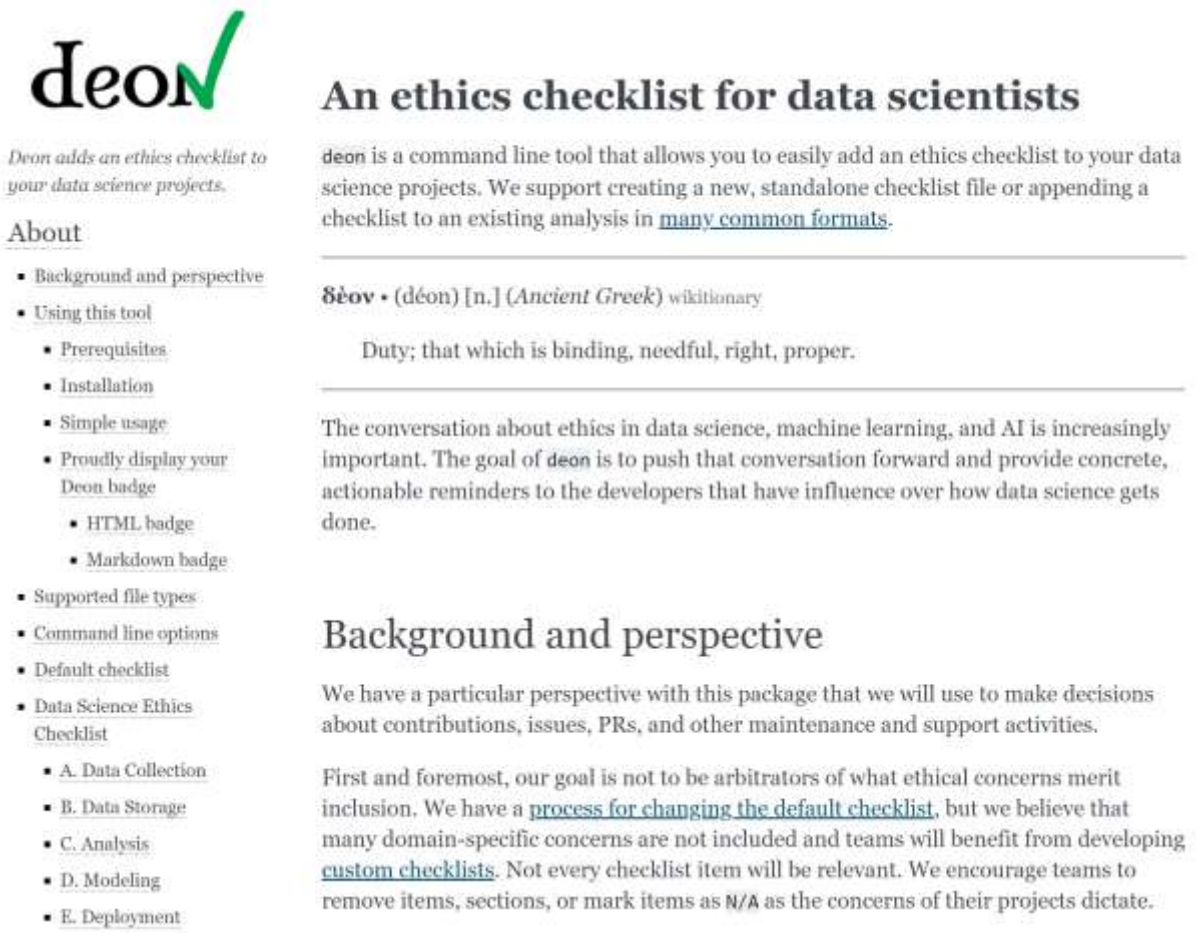
A. Data Collection

B. Data Storage

C. Analysis

D. Modeling

E. Deployment



deon ethics checklist for data scientists:
https://deon.drivendata.org/

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