

Workshop

Ethical Design Matters

Carol J. Smith & Lisa D. Dance
June 2021

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Copyright Statement

Copyright 2021 Carnegie Mellon University and ServiceEase LLC.

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.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon® is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

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

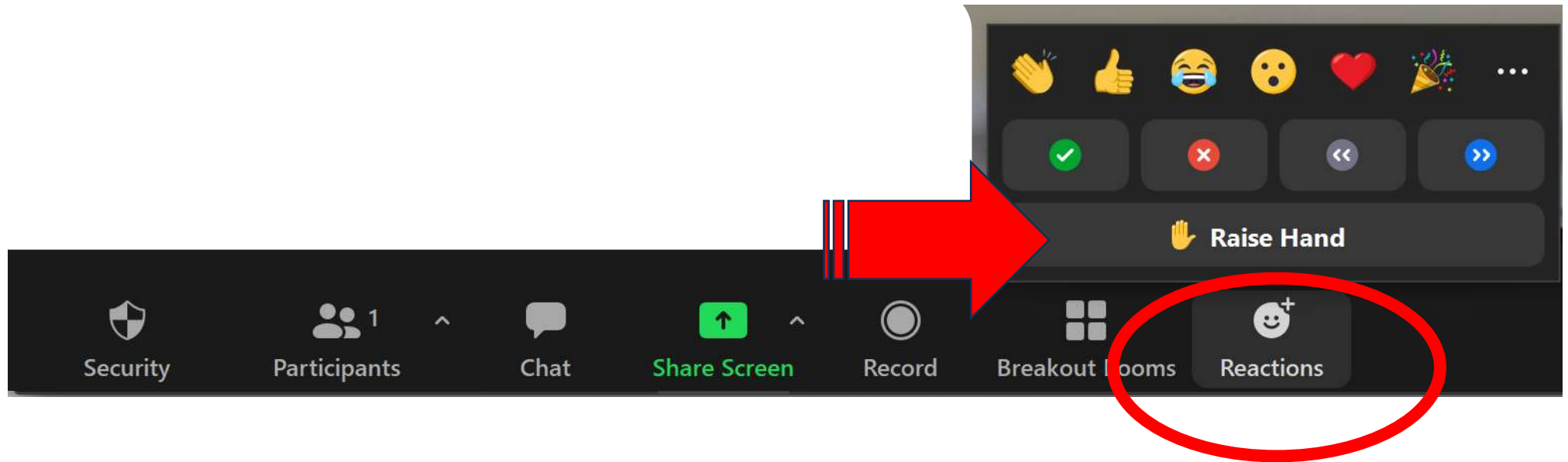
60 min - Activity

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



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.

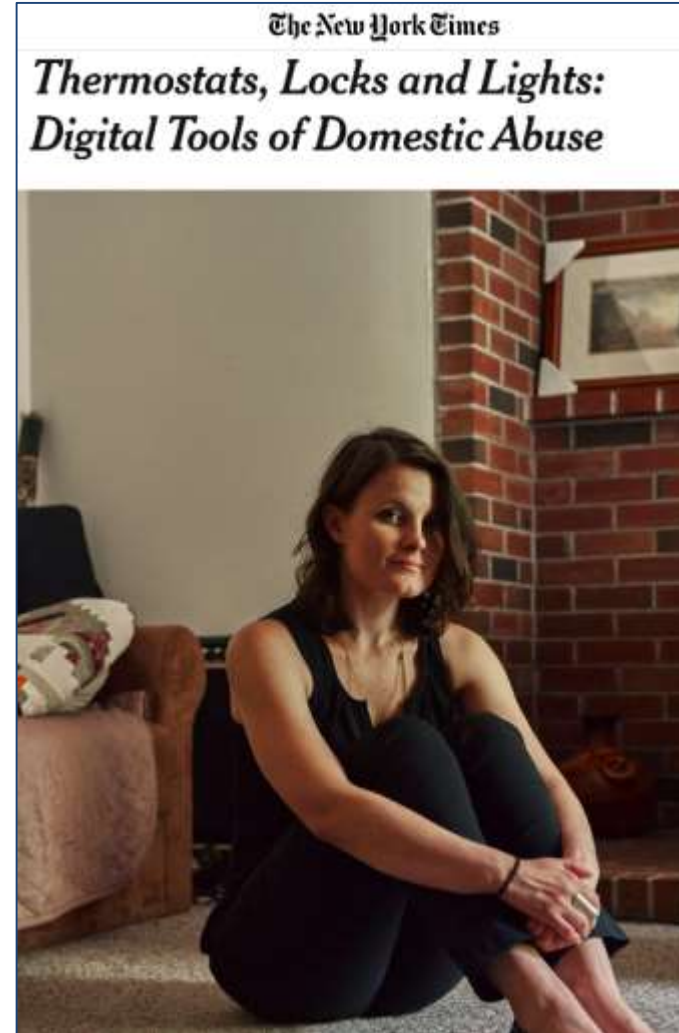
By Mark Haurahan

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.



Assuming That Math Reduces Bias

Vetting applicant resumes



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



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

Ignoring Abusability

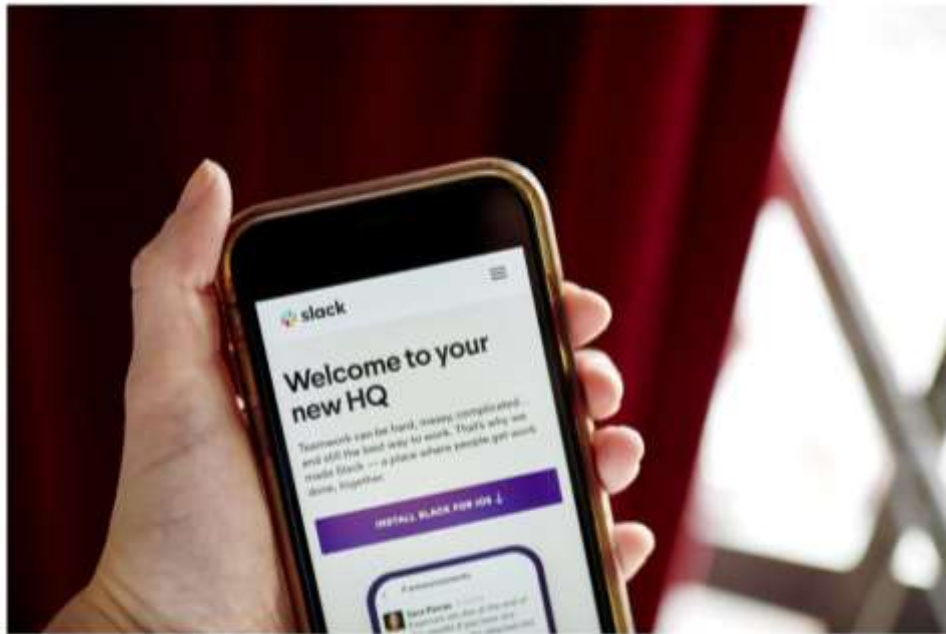
Opting in everyone to messaging

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

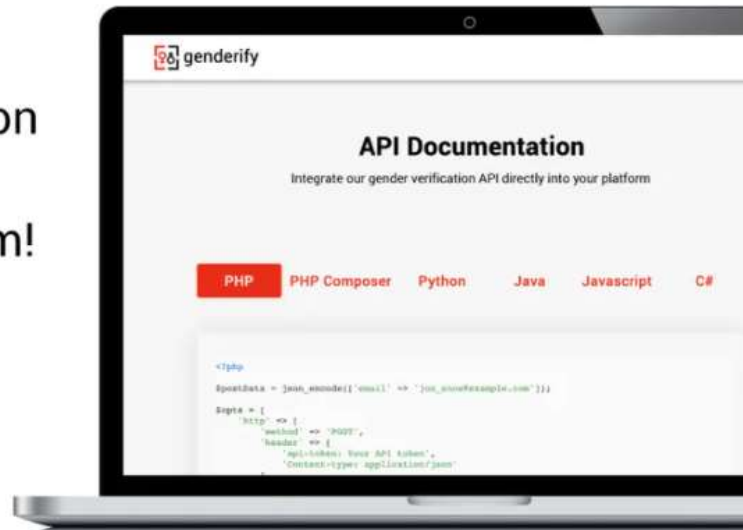
Meghan Smith is a woman, but Dr. Meghan Smith is a man, says Genderify

By [James Vincent](#) | Jul 29, 2020, 6:44am EDT

Integrate our
gender verification
API directly
into your platform!



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

ServiceEase

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

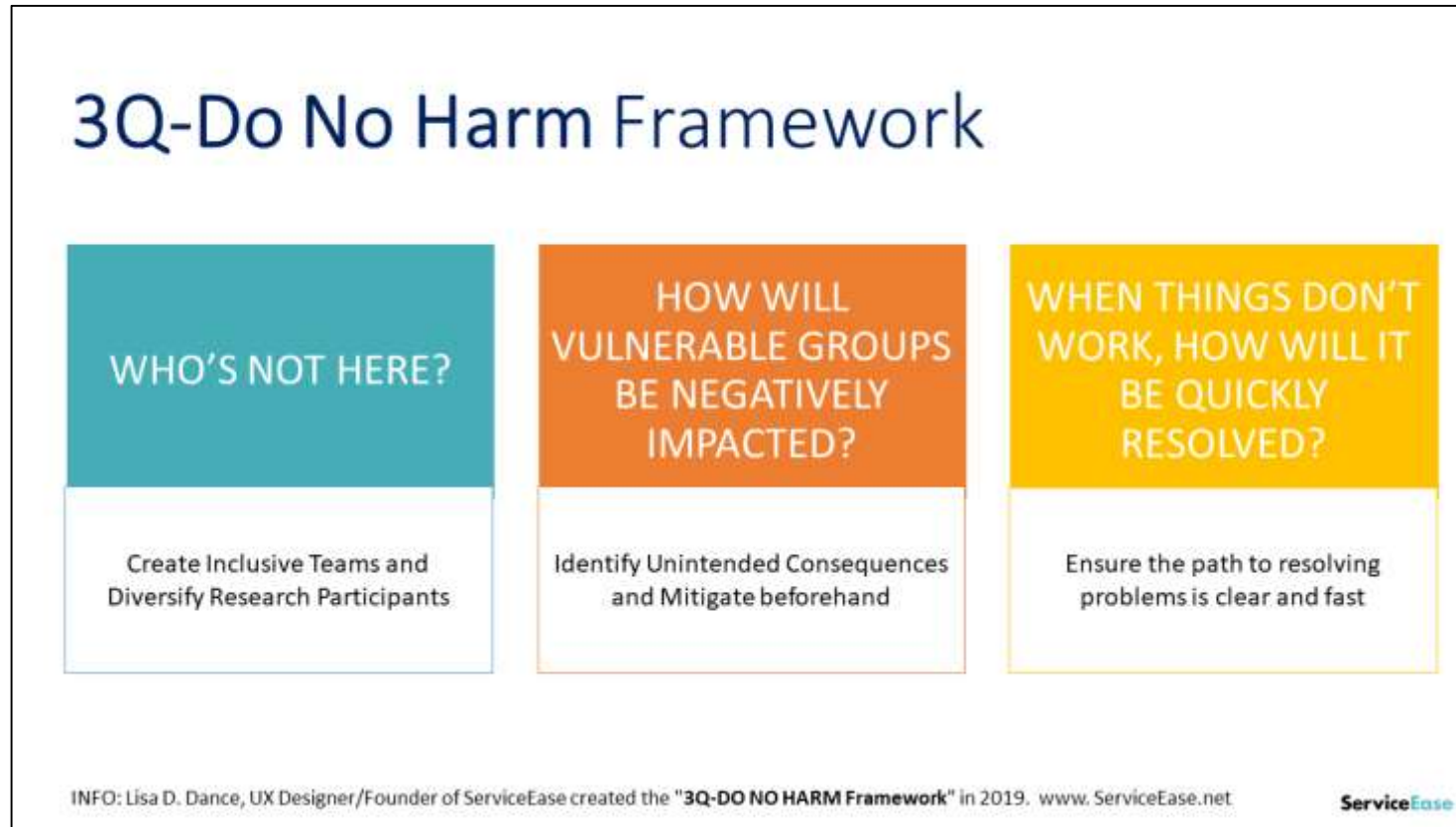
Pair with technical ethics

- 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>.

<p>We will design our AI system with the following in mind:</p> <ul style="list-style-type: none"><input type="checkbox"/> Designated humans have the ultimate responsibility for all decisions and outcomes:<ul style="list-style-type: none">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.<input type="checkbox"/> Significant decisions made by the AI system will be:<ul style="list-style-type: none">explainedable to be overriddenappealable and reversible	<p>We work to speculatively identify the full range of risks and benefits:</p> <ul style="list-style-type: none"><input type="checkbox"/> Harmful, malicious use and consequences, as well as good, beneficial use and consequences.<input type="checkbox"/> We will be cognizant and exhaustively research unintended consequences. <p>We will create plans for the misuse/abuse of the AI system, including the following:</p> <ul style="list-style-type: none"><input type="checkbox"/> communication plans to share pertinent information with all affected people<input type="checkbox"/> mitigation plans for managing the identified speculative risks <p>We value respect and security:</p> <ul style="list-style-type: none"><input type="checkbox"/> incorporating our values of humanity, ethics, equity, fairness, accessibility, diversity, and inclusion<input type="checkbox"/> respecting privacy and data rights (Only necessary data will be collected.)<input type="checkbox"/> providing understandable security methods<input type="checkbox"/> making the AI system robust, valid, and reliable	<p>We value transparency with the goal of engendering trust:</p> <ul style="list-style-type: none"><input type="checkbox"/> The purpose, limitations, and biases of the AI system are explained in plain language.<input type="checkbox"/> Data sources have unambiguous respected sources, and biases are known and explicitly stated.<input type="checkbox"/> Algorithms and models are appropriate and verifiable.<input type="checkbox"/> Confidence and context are presented for humans to base decisions on.<input type="checkbox"/> Transparent justification for recommendations and outcomes is provided.<input type="checkbox"/> Straightforward and interpretable monitoring systems are provided. <p>We value honesty and usability:</p> <ul style="list-style-type: none"><input type="checkbox"/> Humans can easily discern when they are interacting with the AI system vs. a human.<input type="checkbox"/> Humans can easily discern when and why the AI system is taking action and/or making decisions.<input type="checkbox"/> Improvements will be made regularly to meet human needs and technical standards.
---	--	---

Team Signatures and Date:

About the SEI
The Software Engineering Institute is a federally funded research and development center (FEDRC) that works with federal and government organizations, industry, and academia to advance the state of the art in software engineering and cyber security to benefit the public interest. Part of Carnegie Mellon University, the SEI is a national leader in promoting emerging technologies, cybersecurity, software acquisition, and software lifecycle innovation.

Contact Us
CARNegie MELLon UNIVERSITY
SOFTWARE ENGINEERING INSTITUTE
4500 Fifth Avenue, Pittsburgh, PA 15213-2612
WESTMANS
412.268.2000 | 800.261.2470
sei@sei.cmu.edu

©2019 Carnegie Mellon University | SEI | C-1511-2019 | 4/19/2019

Categories of Harm

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

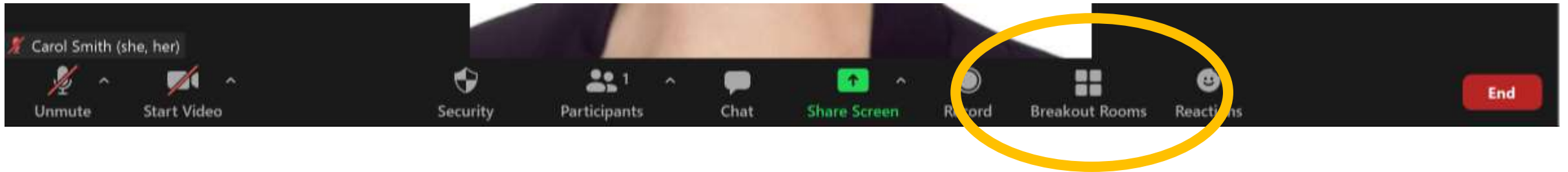
CATEGORIES OF HARM	DEFINITION
FINANCIAL	Negative impact on finances, property, or other resources
HEALTH	Negative impact on mental, emotional, or physical health
TIME	Inefficient or unproductive activities, processes, or systems
FAIRNESS/EQUITY	Perpetuating or facilitating prejudice, bias and/or unfairness
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

ServiceEase

Activity in Mural

Breakout Rooms and Mural

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

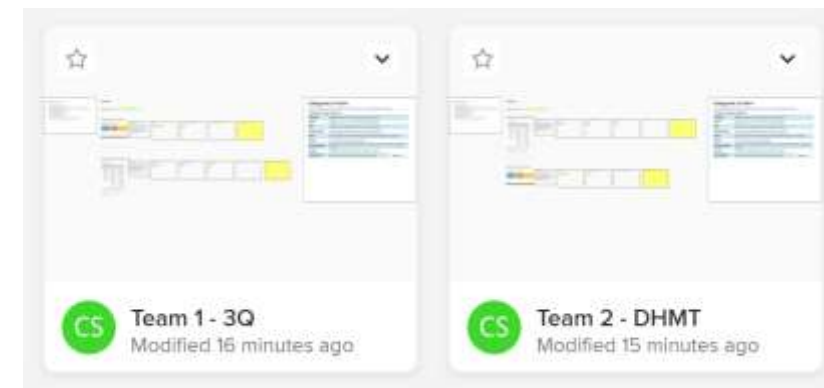


2. Open [Mural](#)

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

3Q-Do No Harm Framework

- MANAGEMENT GOALS**
- HOW WILL STAKEHOLDERS GROUPS BE AFFECTED BY THIS?**
- WHAT CHANGES DOES THIS REQUIRE TO BE SUCCESSFUL?**

Scenario Overview:

A new online content producer, AppFix, will distribute videos, TV shows and other content. They will use information about their users to improve recommendations over time. The system will suggest content and send reminders on what about new content users are interested in based on previous usage. The system will share the primary user's usage with their friends and family. AppFix is planning future videos to limit children's access to content and other safety considerations.

Does that help?



Has our customer group been negatively impacted?



What things don't work, how will it be quality improved?

Top 3 Priorities to address

Take a break as a team between frameworks/checklists

Designing for Ethical AI Checklist

Designing Ethical AI Experiences Checklist and Agreement

As we develop AI products, we will ensure that our AI systems are designed and deployed in a way that respects the rights and privacy of our users and the public good.

- 1. We will ensure that our AI systems are designed and deployed in a way that respects the rights and privacy of our users and the public good.
- 2. We will ensure that our AI systems are designed and deployed in a way that respects the rights and privacy of our users and the public good.
- 3. We will ensure that our AI systems are designed and deployed in a way that respects the rights and privacy of our users and the public good.
- 4. We will ensure that our AI systems are designed and deployed in a way that respects the rights and privacy of our users and the public good.
- 5. We will ensure that our AI systems are designed and deployed in a way that respects the rights and privacy of our users and the public good.
- 6. We will ensure that our AI systems are designed and deployed in a way that respects the rights and privacy of our users and the public good.
- 7. We will ensure that our AI systems are designed and deployed in a way that respects the rights and privacy of our users and the public good.
- 8. We will ensure that our AI systems are designed and deployed in a way that respects the rights and privacy of our users and the public good.
- 9. We will ensure that our AI systems are designed and deployed in a way that respects the rights and privacy of our users and the public good.
- 10. We will ensure that our AI systems are designed and deployed in a way that respects the rights and privacy of our users and the public good.

Scenario Overview:

A new online content producer, AppFix, will produce videos, TV shows and other content. They will use information about their users to improve recommendations over time. The system will suggest content and send reminders on what about new content users are interested in based on previous usage. The system will share the primary user's usage with their friends and family. AppFix is planning future videos to limit children's access to content and other safety considerations.

Accountable to humans



Capable of detecting bias and benefits



Respectful and secure



Honest and usable

Top 3 Priorities to address

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

&

Lisa D. Dance

Twitter: @carologic

Twitter: @ServiceEase

LinkedIn: </in/caroljsmith/>

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 adds an ethics checklist to your data science projects.

About

- [Background and perspective](#)
- [Using this tool](#)
 - [Prerequisites](#)
 - [Installation](#)
 - [Simple usage](#)
- [Proudly display your Deon badge](#)
 - [HTML badge](#)
 - [Markdown badge](#)
- [Supported file types](#)
- [Command line options](#)
- [Default checklist](#)
- [Data Science Ethics Checklist](#)
 - [A. Data Collection](#)
 - [B. Data Storage](#)
 - [C. Analysis](#)
 - [D. Modeling](#)
 - [E. Deployment](#)

An ethics checklist for data scientists

`deon` is a command line tool that allows you to easily add an ethics checklist to your data science projects. We support creating a new, standalone checklist file or appending a checklist to an existing analysis in [many common formats](#).

δῆον • (déon) [n.] (*Ancient Greek*) wikitionary

Duty; that which is binding, needful, right, proper.

The conversation about ethics in data science, machine learning, and AI is increasingly important. The goal of `deon` is to push that conversation forward and provide concrete, actionable reminders to the developers that have influence over how data science gets done.

Background and perspective

We have a particular perspective with this package that we will use to make decisions about contributions, issues, PRs, and other maintenance and support activities.

First and foremost, our goal is not to be arbitrators of what ethical concerns merit inclusion. We have a [process for changing the default checklist](#), but we believe that many domain-specific concerns are not included and teams will benefit from developing [custom checklists](#). Not every checklist item will be relevant. We encourage teams to remove items, sections, or mark items as N/A as the concerns of their projects dictate.

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