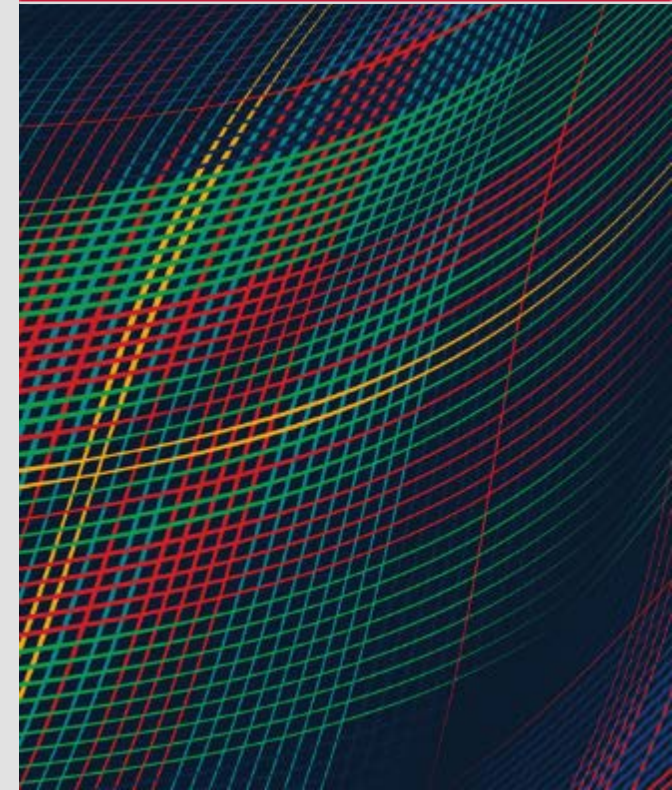


Letting Go of the Numbers: Measuring AI Trustworthiness

ICPRAM 2024

Carol J. Smith

Trust Lab Lead & Principal Research Scientist, AI Division



Copyright Statement

Copyright 2024 Carnegie Mellon University.

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 entity, 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 nor of Carnegie Mellon University - Software Engineering Institute by any such named or represented entity.

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.

DM24-0209

About ACM



ACM, the Association for Computing Machinery (www.acm.org), is the premier global community of computing professionals and students with **nearly 100,000 members in more than 170 countries** interacting with more than 2 million computing professionals worldwide.

OUR MISSION: We help computing professionals to be their best and most creative. We connect them to their peers, to what the latest developments, and **inspire them to advance the profession and make a positive impact on society.**

OUR VISION: We see a world where **computing helps solve tomorrow's problems** – where we use our knowledge and skills to advance the computing profession and make a positive social impact throughout the world.

I am proud to be an ACM Member.

The Distinguished Speakers Program is made possible by



**Association for
Computing Machinery**

Advancing Computing as a Science & Profession

For additional information, please visit <http://dsp.acm.org/>

Human-Machine Teaming



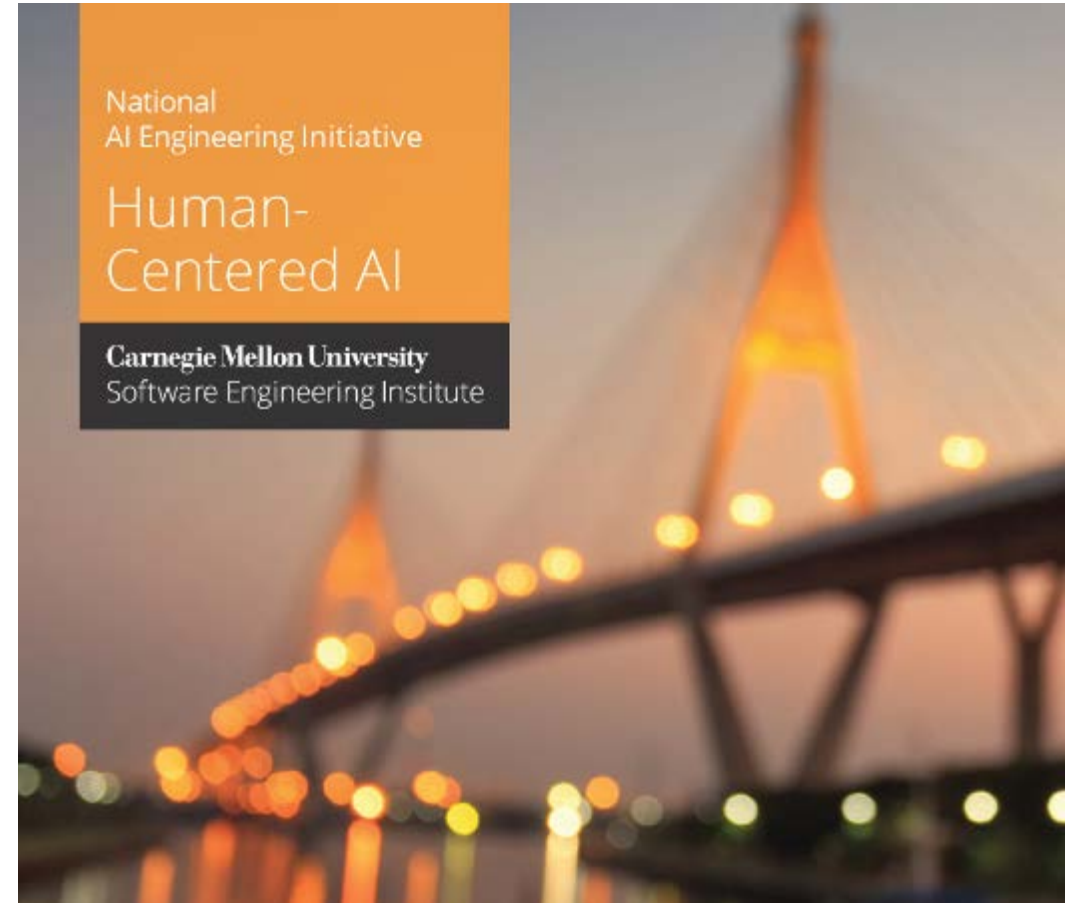
Video: Tesla Autopilot in Heavy LA Traffic by Scott Kubo <https://youtu.be/m3-QzTFxoUg?t=14>

Engineering for Trustworthy AI

Trustworthy AI systems are **designed to work with, and for, people.**

- built for a specific context of use (fit with user needs and tasks)
- with appropriate data, and are
- reliable (robust and secure).

Capabilities are understood, and continuous monitoring and oversight are prioritized.



Human-Centered AI, Software Engineering Institute:
<https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362>

AI must be designed to work with, and for, people. Trustworthy, human-centered, and responsible.



How do we Measure Trustworthiness?



Can we accurately predict the future?



Bernard Parker, left, was rated high risk; Dylan Pugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Can we use data to reduce bias in systematically prejudiced organizations?



Can Anyone?



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.

Amazon scraps secret AI recruiting tool that showed bias against women
By Jeffrey Dastin. October 9, 2018. Reuters.

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

Nearly three years after the company was called out, it hasn't gone beyond a quick workaround

By James Vincent | Jan 12, 2018, 10:35am EST



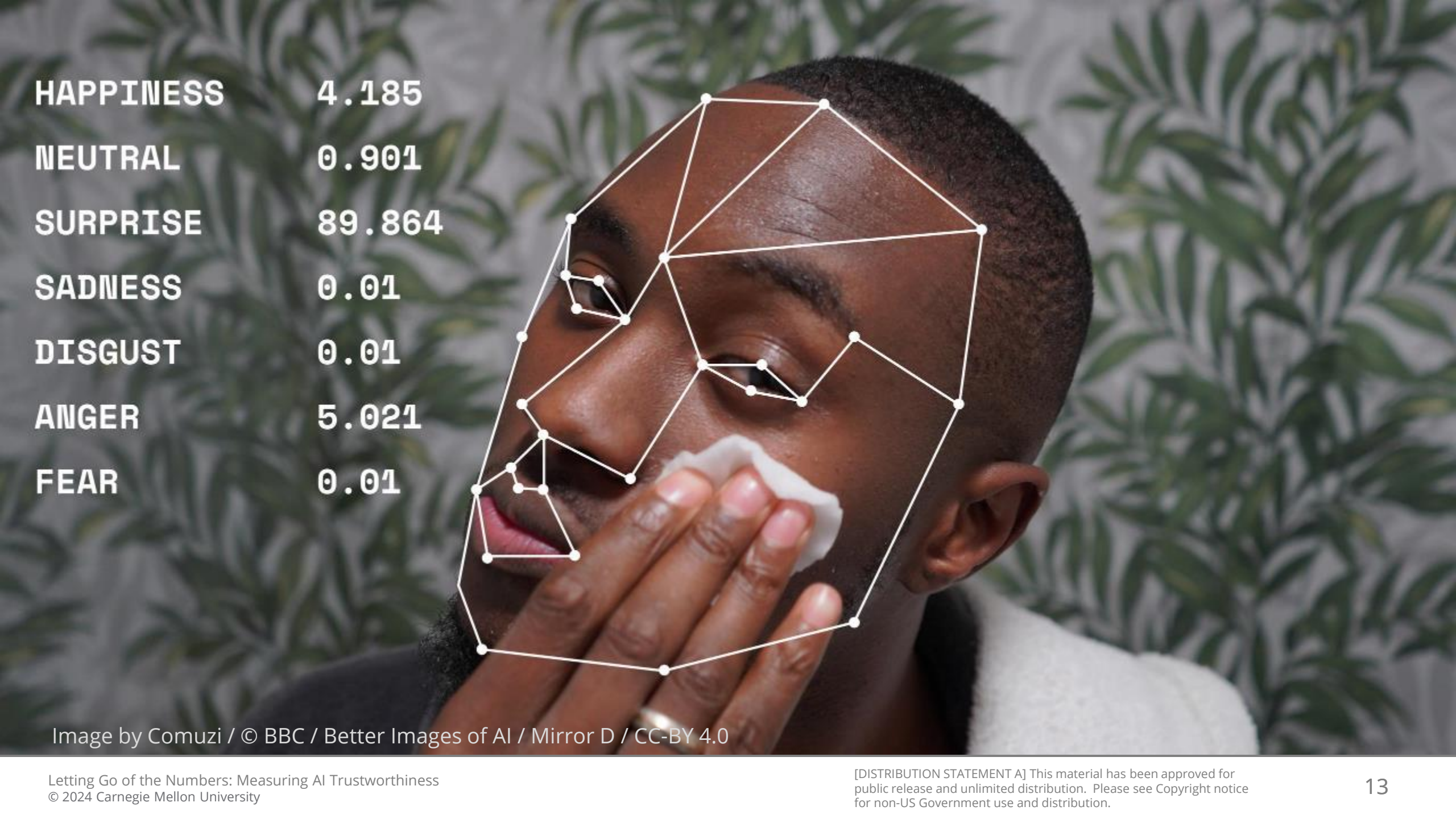
The AI algorithms in Google Photos sort images by a number of categories. | Photo by Vjeran Pavic / The Verge

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech. By James Vincent Jan 12, 2018. The Verge.

How about generative AI or LLMs?



Jason Allen's A.I.-generated work, "Théâtre D'opéra Spatial," took first place in the digital category at the Colorado State Fair. Credit... via Jason Allen. <https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html>



HAPPINESS	4.185
NEUTRAL	0.901
SURPRISE	89.864
SADNESS	0.01
DISGUST	0.01
ANGER	5.021
FEAR	0.01

Image by Comuzi / © BBC / Better Images of AI / Mirror D / CC-BY 4.0

Humans create and use imperfect machines.



Quant Performance Evaluations are Necessary

- Evaluate accuracy, precision, recall
- Ensure it is robust, secure, reliable
- Speed of system
- Scalability

Relative simplicity of these methods is appealing,
but **these are not adequate.**

Overnight Flight from US to Rome

Quantitative

- Plane arrived 20+ min. early.
- Reduced fuel use.
- Reduced emissions.

Qualitative

- Delayed meal delivery.
- Reduced sleep time.
- An uncomfortable night.



Overnight Flight from US to Rome

Quantitative

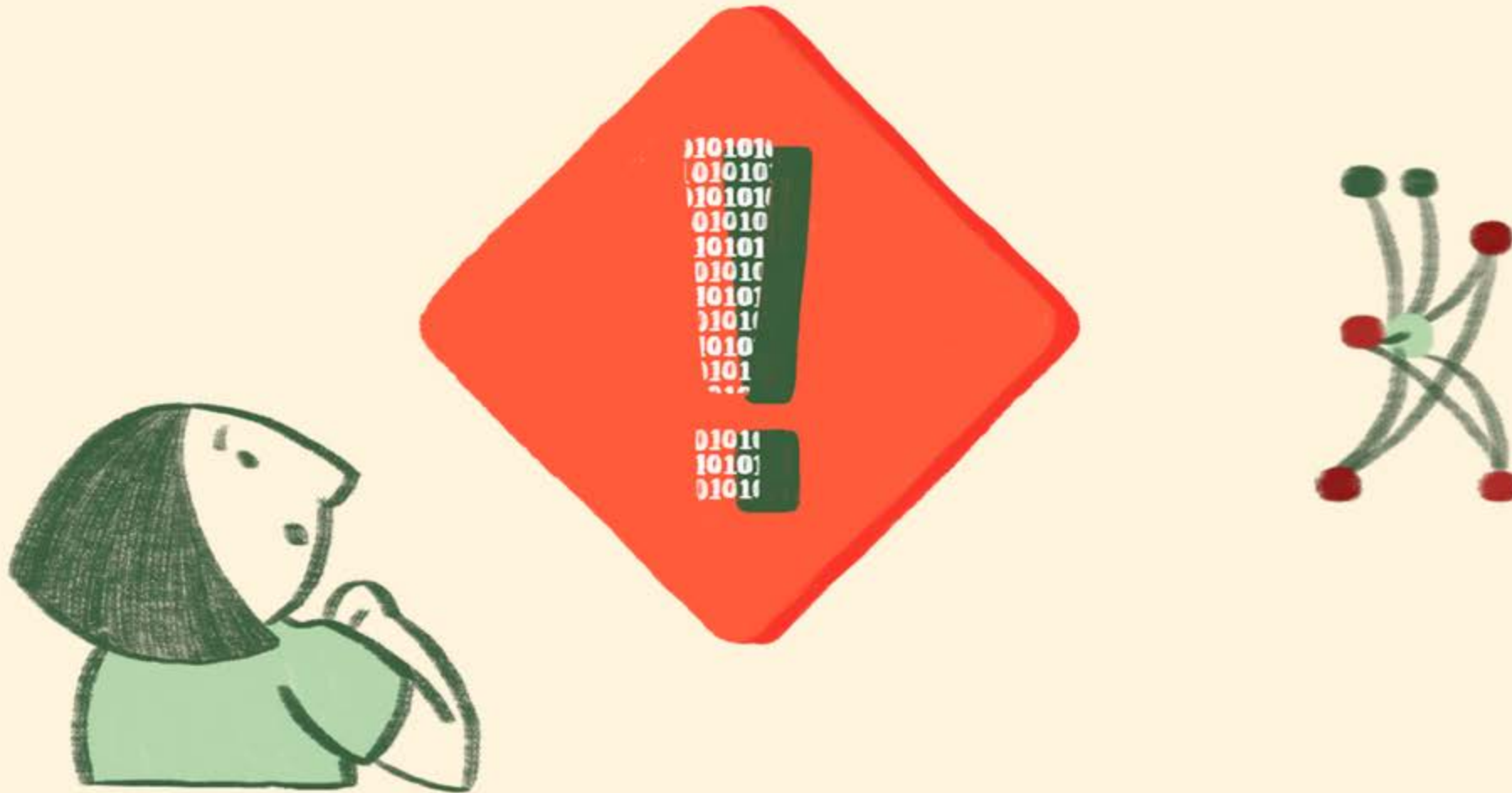
- Plane arrived 20+ min. early.
- Reduced fuel use.
- Reduced emissions.

Qualitative

- Delayed meal delivery.
- Reduced sleep time.
- An uncomfortable night.



Trustworthiness Requires Qualitative Measures



All systems will have some form of bias

Complete objectivity is misleading.

Bias can have purpose and can be helpful.

Bias contributes to and is emphasized by decisions.

We must ensure we

- identify and understand bias
- reduce unintended and/or harmful bias.

Risks due to bias are lower when no information about people is present.

Bias Due to Data, Algorithm, and Training



Photo by sunlightfoundation
<https://www.flickr.com/photos/sunlightfoundation/2385174105>

“Data is a function of our history...
The past dwells within...
Showing us the inequalities
that have always been there.”

Joy Buolamwini, Algorithmic Justice League
Coded Gaze
Movie: Coded Bias on Netflix

Photo: Joy Buolamwini on The Open Mind: Algorithmic Justice.
Jan 12, 2019. <https://www.youtube.com/watch?v=hwHnXdoSSFY>

THE
OPEN MIND



Start with Data

Need to confirm:

- Appropriateness
- Provenance and understanding of data composition and variance
- Information about people and potential risks

Translated Wikipedia Biographies

The Translated Wikipedia Biographies dataset has been designed to evaluate gender accuracy in long text translations (multiple sentences or passages). The set has been designed to analyze common gender errors in machine translation like incorrect gender choices in anaphora resolutions, possessives and gender agreement.

PUBLISHERS: Google LLC	INDUSTRY TYPE: Corporate - Tech	DATASET AUTHORS: Anja Axelsen, Google Michelle Lynch, Google Kanna Stella, Google Rishi Mishra, Google
FUNDERS: Google LLC	FUNDING TYPE: Private Funding	DATASET CONTACT: transLata-gender@ml.google.com
OUTPUT PURPOSES: Testing	KEY APPLICATIONS: Machine Translation: Gender Accuracy PRIMARY MOTIVATION(S): Study gender accuracy in translations beyond the sentence or paragraph level and to explore their utility for various research.	INTENDED END-USER SYSTEMS OR USE CASE(S): To evaluate gender accuracy on translations beyond the sentence (multiple sentences or passages). The set is focused on the presence of the specific linguistic phenomena to evaluate the most common contextual errors. • Spanish to English: masculine • Spanish to English: Neutral to gender-specific pronouns • English to Spanish, German: Gender agreement
PRIMARY DATA TYPE(S): Non-Sensitive Public Data about people	DATASET STATISTICS: Total instances: 100 Masculine biographies (sentences): 33 Masculine biographies (paragraphs): 31 Feminine biographies (sentences): 33 Feminine biographies (paragraphs): 31 Both female & sport teams (sentences): 11 Both female & sport teams (paragraphs): 11	DESCRIPTION OF DATASET: This dataset is based on publicly available data on public and/or historical figures (Wikipedia articles) of a given language in 1995. The dataset has 100 instances and each instance contains the first 10 to 15 sentences from a Wikipedia article. Articles are written in native English and have been professionally translated to Spanish and German. 150 of these instances represent a person with an associated stated gender and 12 are related with rock bands or sport teams (unassociated genderless).
	DATASET SOURCES: • Source text: English Wikipedia • Target Text: Professional translators	HOW TO INTERPRET A DATASET POINT: Each datapoint refers to a central entry that can be a person, listed as feminine or masculine, a rock band or a sport team (genderless). Each entry is represented by a long text translation (multiple sentences, sentences or paragraphs) belonging to that main entry.
PRIMARY DATA MODALITY: Textual Data	EXAMPLE OF ACTUAL DATA POINT WITH DESCRIPTIONS:	
	sourceLanguage: es	Language of the original text.
	targetLanguage: es	Language of the translation.
	documentID: 1	ID generated to identify all the sentences belonging to the same passage.
	stringID: 1-1	Generated by the documentID and sentence number in the passage.
	sourceText: "Elisa-Laura Wittchen (nacida el 25 de junio de 1982) es una cantante alemana, que actualmente compete por el Campeonato Mundial de Fútbol."	Text from Wikipedia in source language (punctuations and quotes removed).
	translatedText: "Elisa-Laura Wittchen (nacida el 25 de junio de 1982) es una cantante alemana, que actualmente compete por el Campeonato Mundial de Fútbol."	Translation of the Wikipedia source text into the target text.
	genderInSource: Female	Identified as Female, Male, Neutral.
	stringName: Elisa-Laura Wittchen	Name of the main entry according Wikipedia.
	sourceURL: https://es.wikipedia.org/wiki/Elisa-Laura_Wittchen#/media/Elisa-Laura_Wittchen	Link to the Wikipedia article at the time of extraction. Please consider that content in Wikipedia articles can be modified so differences may be found if the article has been re-edited.

Data Card v1.0 • Published June 2021 • Updated Sep 2021 Page 1 of 3

Sample data card, source: [Pushkarna et al., 2022](#)

Spotted Lanternfly



Spotted lanternfly Life Cycle. Published by Oxford University Press on behalf of Entomological Society of America 2021., Public domain, via Wikimedia Commons.

Data Provenance

- Researcher's motivation
- Collection process
- Data included and excluded
 - Which stage of life cycle?
 - Locations?
- Recommended uses, etc.
- Historical patterns of negative bias
- Sensitivity of data



Spotted lanternfly displaying underwing.
WanderingMogwai, CC BY-SA, via Wikimedia Commons

Identification of Inherent Bias

Understand inherent bias and amount of variance in dataset due to data provenance.

Bias can be both purposeful and unintended influences

- **Purposeful:** provenance of data, collection process, etc.
- **Unintended:** existing systematic bias that may or may not be known or is only revealed as the system is developed.



All systems are biased.

**Each decision
creates and affects bias.**

**Bias can have purpose
and can be helpful.**

Unwanted bias can lead to inequitable outcomes

At the surface, AI systems can seem objective and impartial.

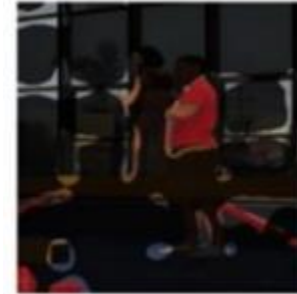
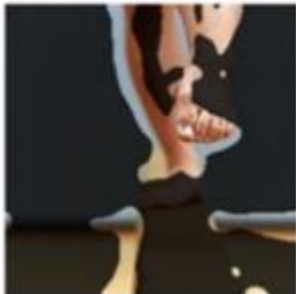
Digging deeper reveals that AI systems can reinforce discrimination against historically marginalized groups

- Alignment problems
- Scale problems
- Multiplicity problems

Fairness research led by Anusha Sinha, AI Division at the SEI

Bias can result in the right decision for the wrong reasons

Images correctly classified as "balance beam"



Original images sourced from ImageNet

Fairness research led by Anusha Sinha, AI Division at the SEI

Wrong reasons can lead to poor real-world performance

A low-stakes example:



Source: ImageNet

Ground truth: horizontal bar

Predicted: balance beam

A high-stakes example:



Ground truth: Carol in town during protest

Predicted: Carol organized protest

Fairness research led by Anusha Sinha, AI Division at the SEI

Mitigation of Bias is Complex

Removing all bias is impossible

- Removing obvious indicators (gender, zip code, etc.) reduces the ability to track bias.
- Invisible indicators are concealed in the data.
- Share awareness of bias for all audiences (developers, purchasers, users).

Getting to Trustworthiness



An AI system's potential is bound to stakeholders' perceptions of its trustworthiness



Capitalize on Human Strengths

Humans are (still) better
at many activities:

Exposing Bias

Identifying downstream impacts

Judgment

Recognizing Bias

Responding to change

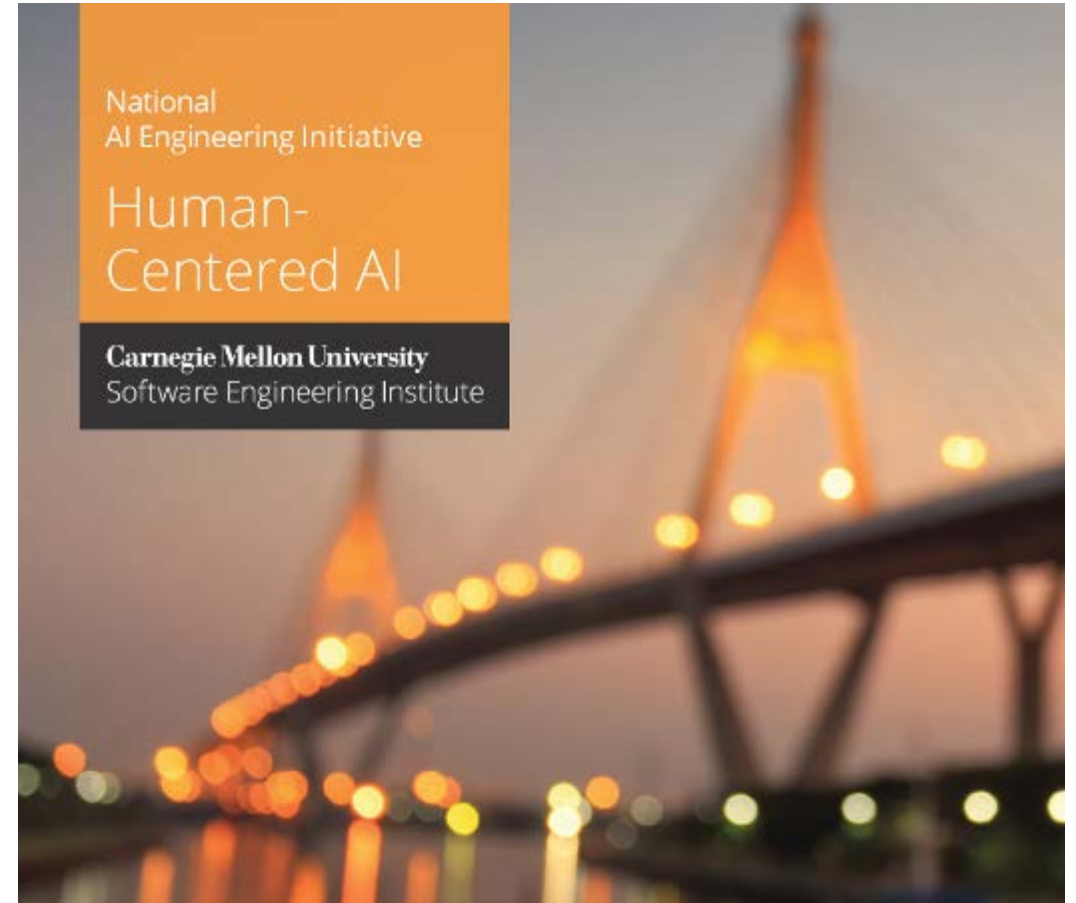
Socio-political nuance

Taking context into consideration

Amanda Muller and Carol Smith. 2022. Perceptions of Function Allocation between Humans and AI-Enabled Systems. UXPA 2022 (pre-print).
<https://uxpa2022.org/sessions/perceptions-of-function-allocation-between-humans-and-ai-enabled-systems/>

Trust Should Not be the Default

- Dynamic systems
Data drift, poisoning, system failures
- Dynamic contexts
Weather, adversaries
- Human judgement
Intuition, situational awareness,
fatigue



Human-Centered AI, Software Engineering Institute:
<https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=735362>

Trust is Contextual

Trust is personal - a dynamic psychological state.
We calibrate trust based on personal experiences, current context,
and available evidence of 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.

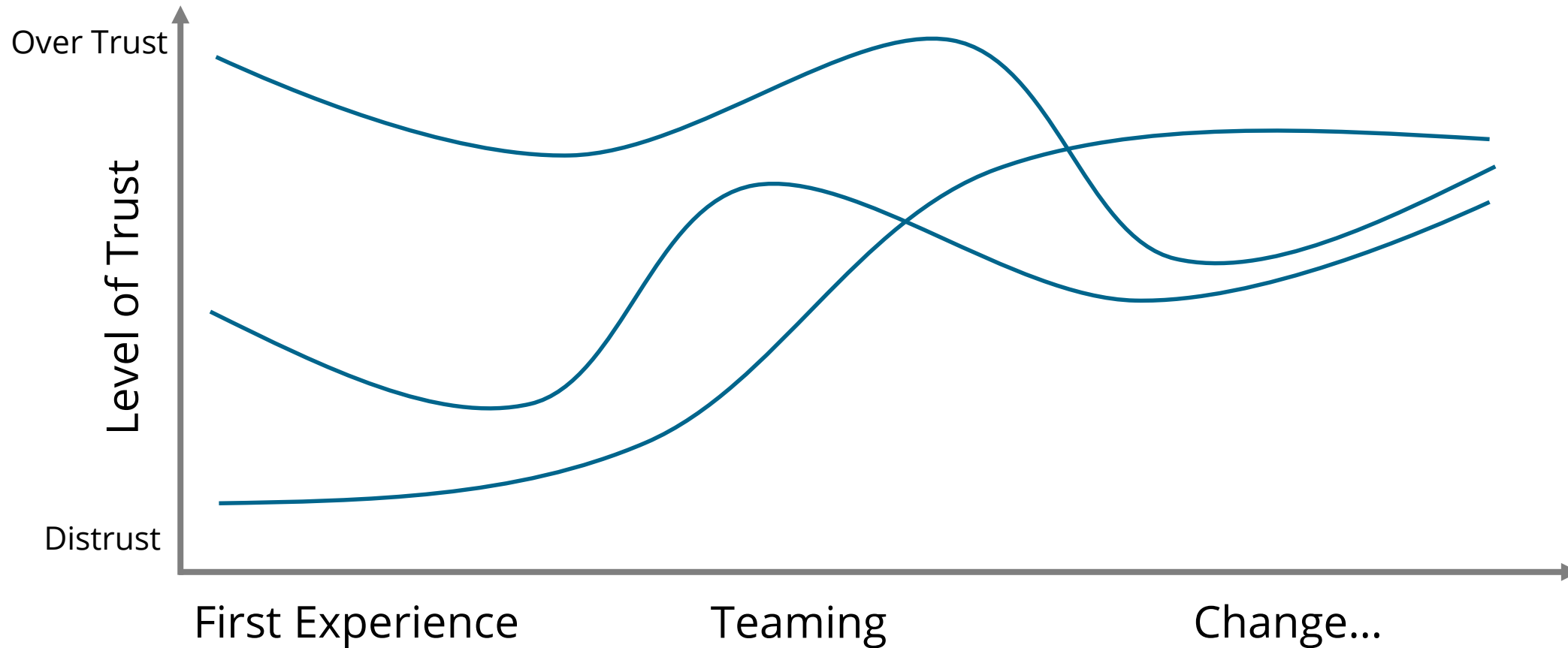
Over Trust

Trust exceeding
system capabilities -
may lead to misuse.



John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. Hum Factors 46, 1 (March 2004), 50–80. DOI:https://doi.org/10.1518/hfes.46.1.50_30392
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>

Trust is Complex and Transient



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>

Design for Use, Context, Trustworthiness



Who will use it?

Scientist?

Grocer?

Understand Stakeholders

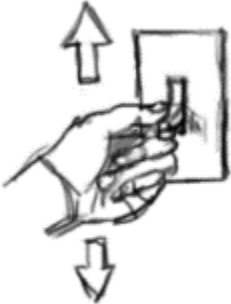
- Who will use the system?
- How well are current systems accepted?
- What are the existing issues?
- When and in what context?

Trustworthy Systems

- Uphold Responsible AI principles
- Utilize data appropriate for task
- Designed for the human-machine team to complete their mission
- Augment human teammates and meet their needs (human-centered)
- Consistently provide adequate evidence of current capabilities and integrity in the current context.

Measurements of Trustworthiness

Usability



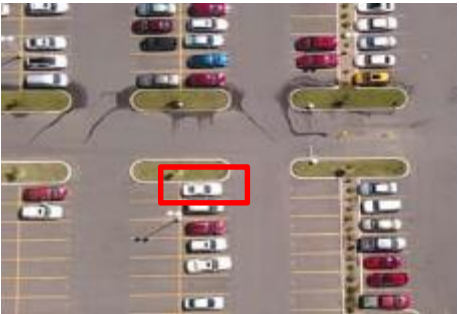
Explainability



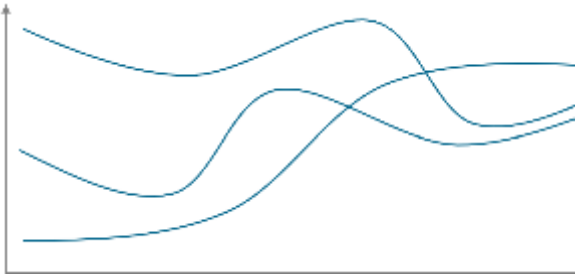
Fairness



Likelihood of Failure



Usage





How is the AI system making decisions?
Explainability

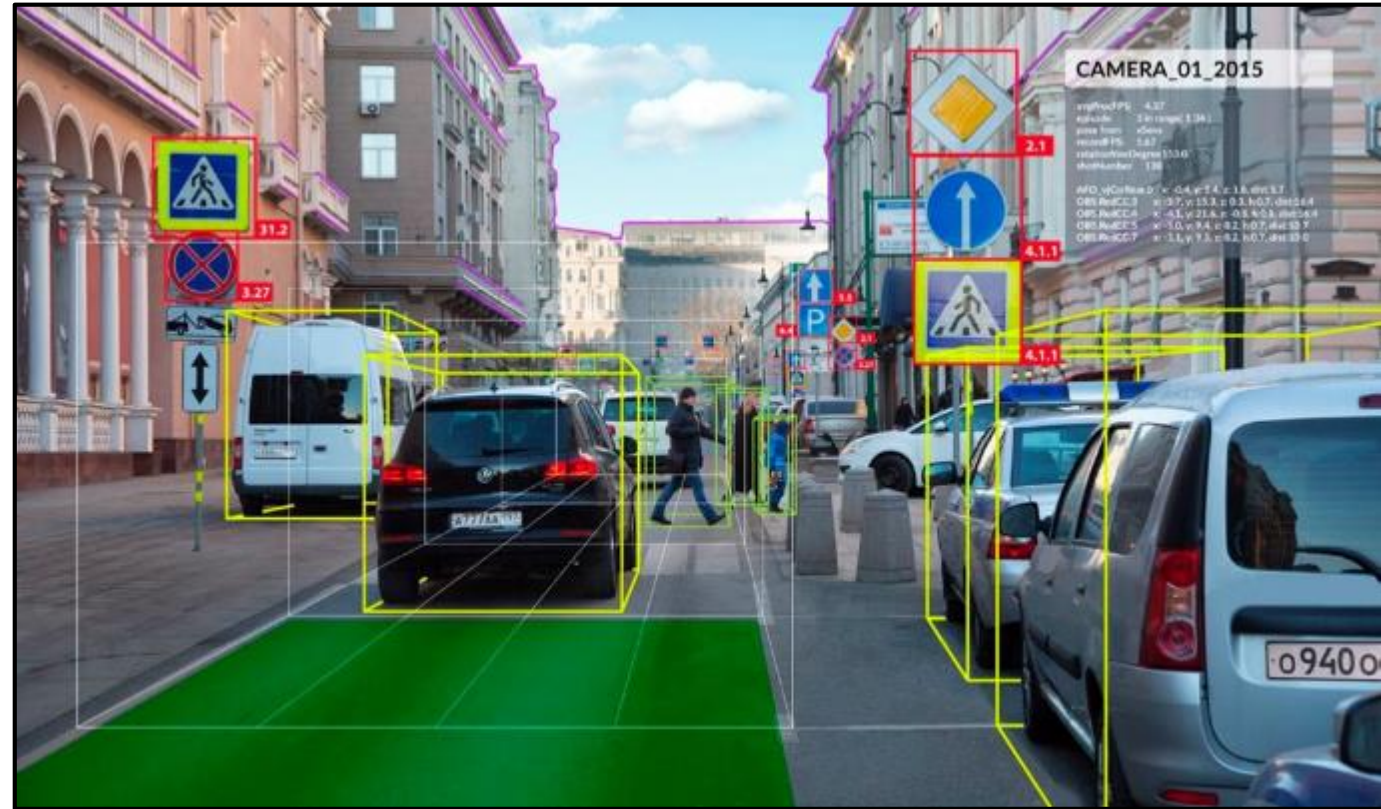
Photo GOODFELLOW AIR FORCE BASE, TX, UNITED STATES, 11.05.2020 by Airman 1st Class Ethan Sherwood, 17th Training Wing Public Affairs <https://www.dvidshub.net/image/6443325/drones-goodfellow>

Explainability reveals decision-making processes

Interpretability facilitates optimization and evaluation

([Doshi-Velez & Kim, 2017](#))

- Safety
- Ethics
- Mismatched objectives
- Multi-objective tradeoffs



Example of a computer vision system, Source: [Welker Media](#)

Explainability research led by Violet Turri, AI Division Trust Lab at the SEI

Explanations can Illuminate Unintended System Behavior

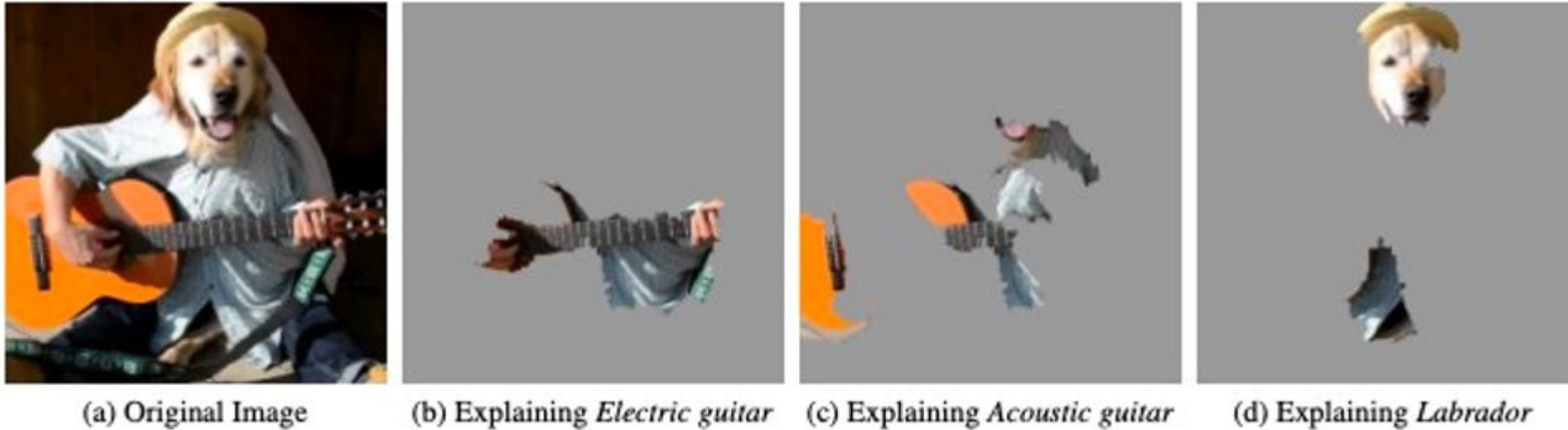


Figure 4: Explaining an image classification prediction made by Google’s Inception network, highlighting positive pixels. The top 3 classes predicted are “Electric Guitar” ($p = 0.32$), “Acoustic guitar” ($p = 0.24$) and “Labrador” ($p = 0.21$)

Sample explanation of an image classifier, Source: [Explainable AI: current status and future directions](#)

Explainability research led by Violet Turri, AI Division Trust Lab at the SEI

A photograph of a modern building at night. The building has a dark facade with many windows, some of which are illuminated from within. In the foreground, there is a walkway with a colorful, multi-colored light display that looks like a rainbow or a spectrum of colors. The sky is dark blue, suggesting dusk or night. The overall scene is a mix of modern architecture and vibrant lighting.

Understanding the Likelihood of Failure

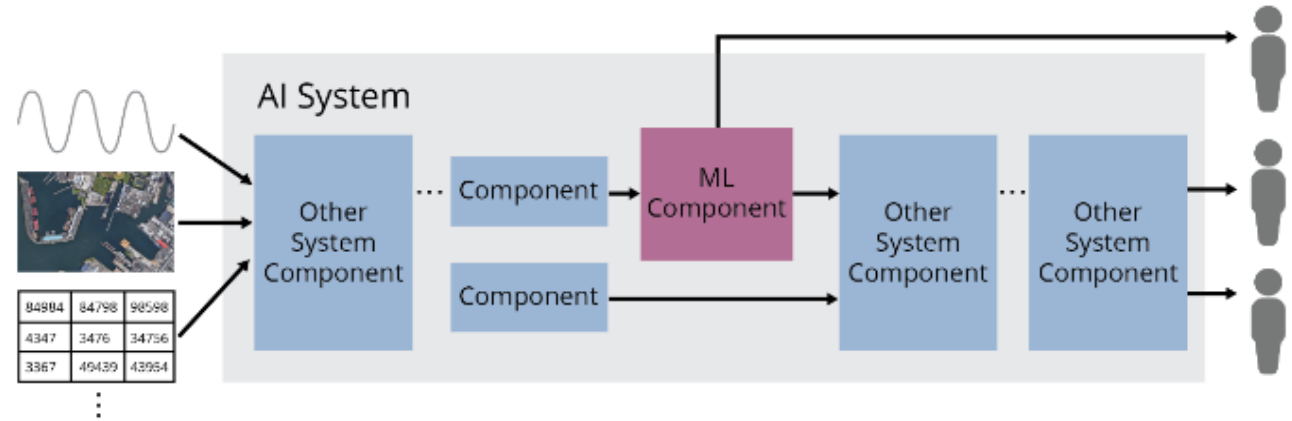
Accurate confidence measures can inform better decision making in complex contexts

People need

- situational awareness (system and context), and
- probability of failure.

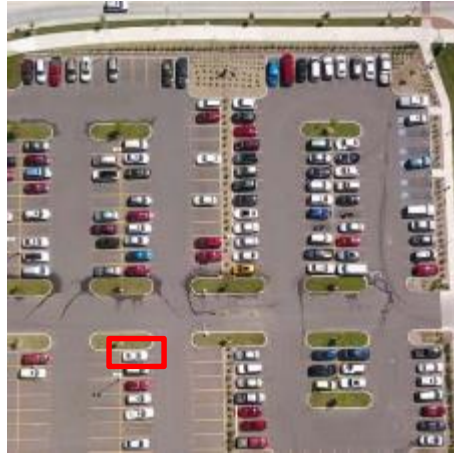
Decide what to do next

- inform other parts of the system
- alert an analyst, use another sensor, etc.



Uncertainty Quantification research led by Eric Heim, AI Division at the SEI

Why is confidence important?



0.2203 Confident



0.9637 Confident

- More informed decision making and prioritization
- Focus on the car on the right
- Use additional resources to confirm

Uncertainty Quantification research led by Eric Heim, AI Division at the SEI

Why are context and shifting environments important?

Many training data sets do not provide sufficient coverage of cases that can be encountered in the deployment environment.

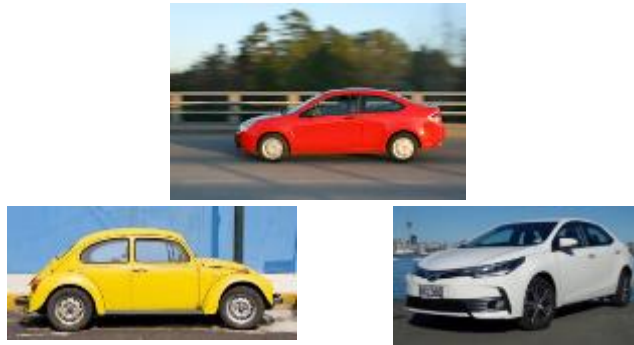


Uncertainty Quantification research led by Eric Heim, AI Division at the SEI

Why are context and shifting environments important?

Many training data sets do not provide sufficient coverage of cases that can be encountered in the deployment environment.

Train Set



Encountered During Deployment



Potential causes of shifts:

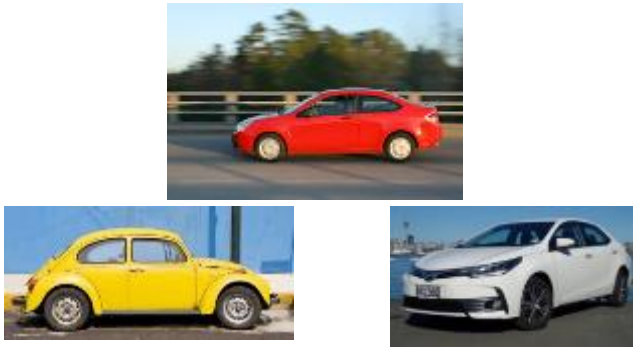
- Sensor failure or degradation

Uncertainty Quantification research led by Eric Heim, AI Division at the SEI

Why are context and shifting environments important?

Many training data sets do not provide sufficient coverage of cases that can be encountered in the deployment environment.

Train Set



Encountered During Deployment



Potential causes of shifts:

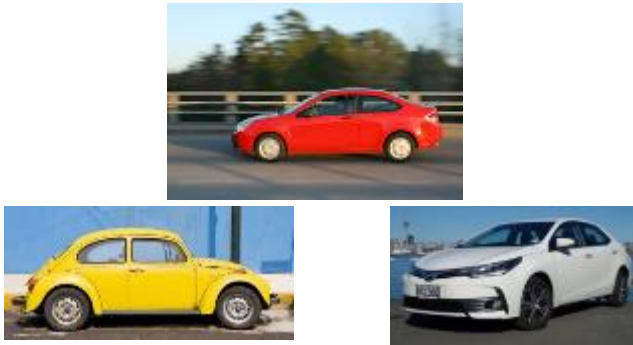
- Sensor failure or degradation
- Unidentified biases

Uncertainty Quantification research led by Eric Heim, AI Division at the SEI

Why are context and shifting environments important?

Many training data sets do not provide sufficient coverage of cases that can be encountered in the deployment environment.

Train Set



Encountered During Deployment



Potential causes of shifts:

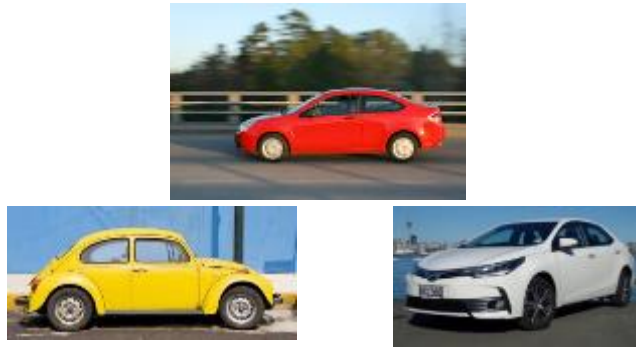
- Sensor failure or degradation
- Unidentified biases
- Unaccounted for changes in data pipelines

Uncertainty Quantification research led by Eric Heim, AI Division at the SEI

Why are context and shifting environments important?

Many training data sets do not provide sufficient coverage of cases that can be encountered in the deployment environment.

Train Set



Encountered During Deployment



Potential causes of shifts:

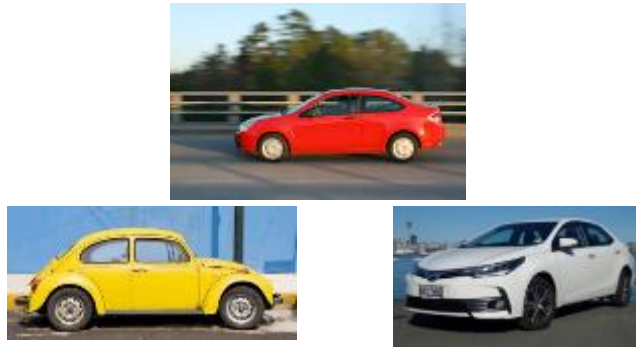
- Sensor failure or degradation
- Unidentified biases
- Unaccounted for changes in data pipelines
- Change in context

Uncertainty Quantification research led by Eric Heim, AI Division at the SEI

Why are context and shifting environments important?

Many training data sets do not provide sufficient coverage of cases that can be encountered in the deployment environment.

Train Set



Encountered During Deployment



Potential causes of shifts:

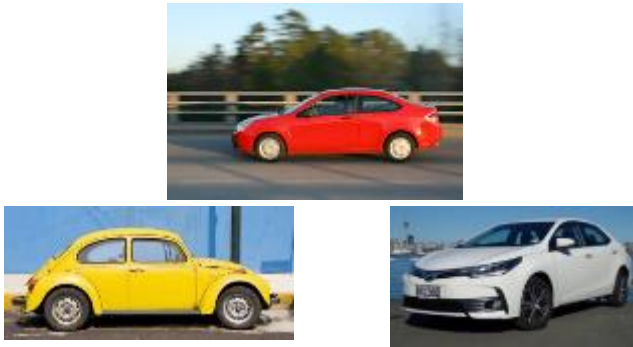
- Sensor failure or degradation
- Unidentified biases
- Unaccounted for changes in data pipelines
- Change in context
- Rare, but possible cases

Uncertainty Quantification research led by Eric Heim, AI Division at the SEI

Why are context and shifting environments important?

Many training data sets do not provide sufficient coverage of cases that can be encountered in the deployment environment.

Train Set



Encountered During Deployment



Potential causes of shifts:

- Sensor failure or degradation
- Unidentified biases
- Unaccounted for changes in data pipelines
- Change in context
- Rare, but possible cases
- **Novel, but relevant classes**

Uncertainty Quantification research led by Eric Heim, AI Division at the SEI

Why is the model uncertain?

- What is the cause?
- How do we want the system to respond when it encounters new information – new situations?
- What is the appropriate way to communicate the likelihood of failure?

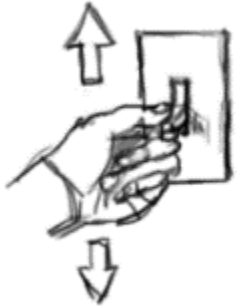


Testing for Trustworthiness



Measurements of Trustworthiness

Usability



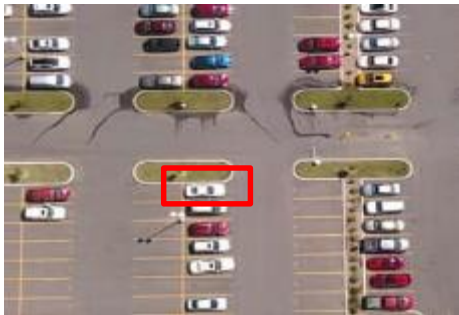
Explainability



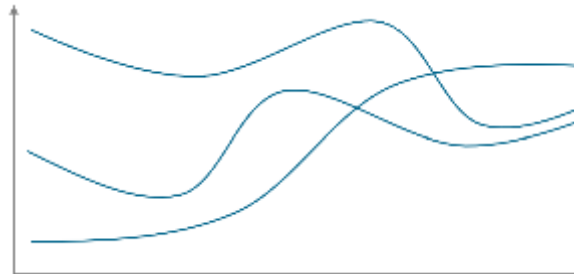
Fairness



Likelihood of Failure



Usage



Conversations for Understanding

Difficult Topics

- What do we value?
- Who could be hurt?
- What lines won't our AI 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 On Unsplash
https://unsplash.com/@msgrace?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText





Speculation Keeps People Safe

Activate Curiosity

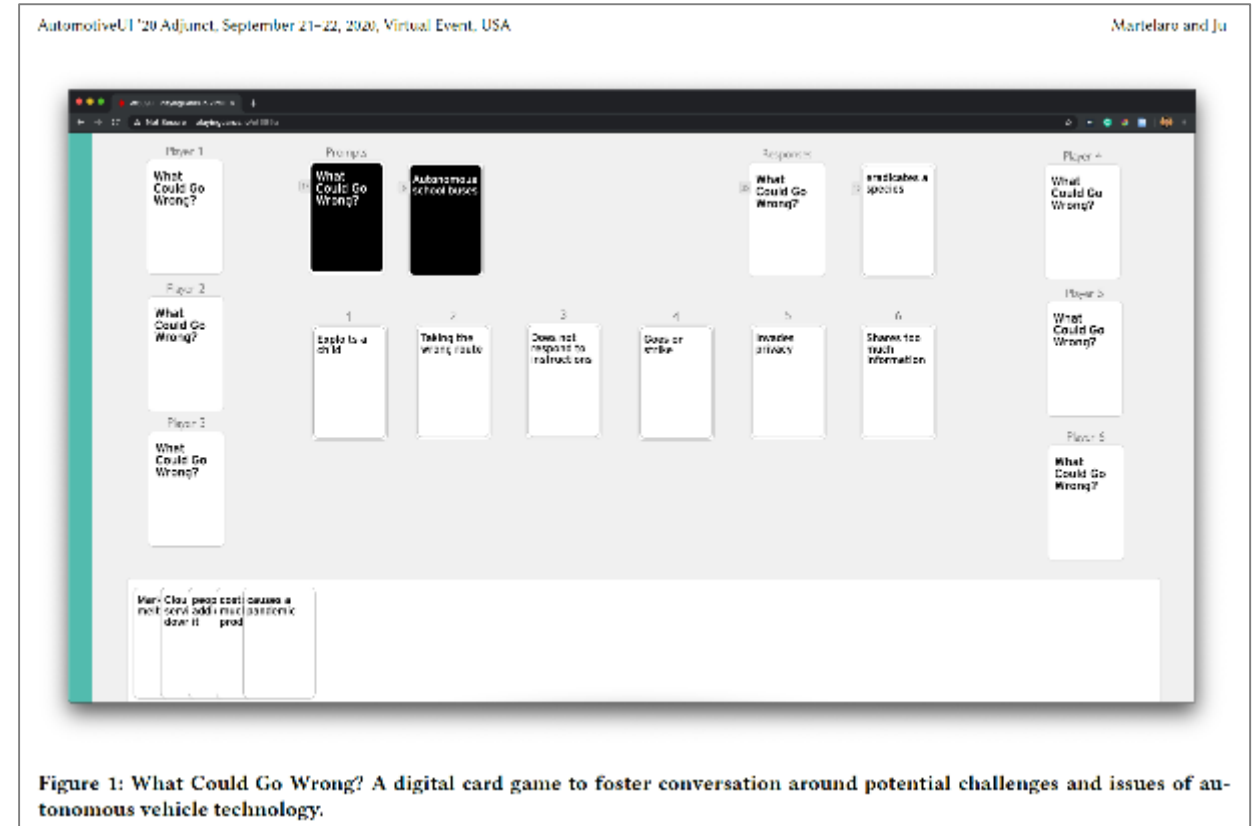
Speculate about misuse and abuse

- Unintended and unwanted consequences
- Negative consequences for people who are frequently marginalized

Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020.

Card Game: What Could Go Wrong?

Foster conversations around potential challenges and issues with complex technologies.



Nikolas Martelaro and Wendy Ju. 2020. What Could Go Wrong? Exploring the Downsides of Autonomous Vehicles. In 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '20). Association for Computing Machinery, New York, NY, USA, 99-101. <https://doi.org/10.1145/3409251.3411734>

Abusability Testing

1) Value proposition

Benefits tech brings to individuals, society

2) Vulnerabilities

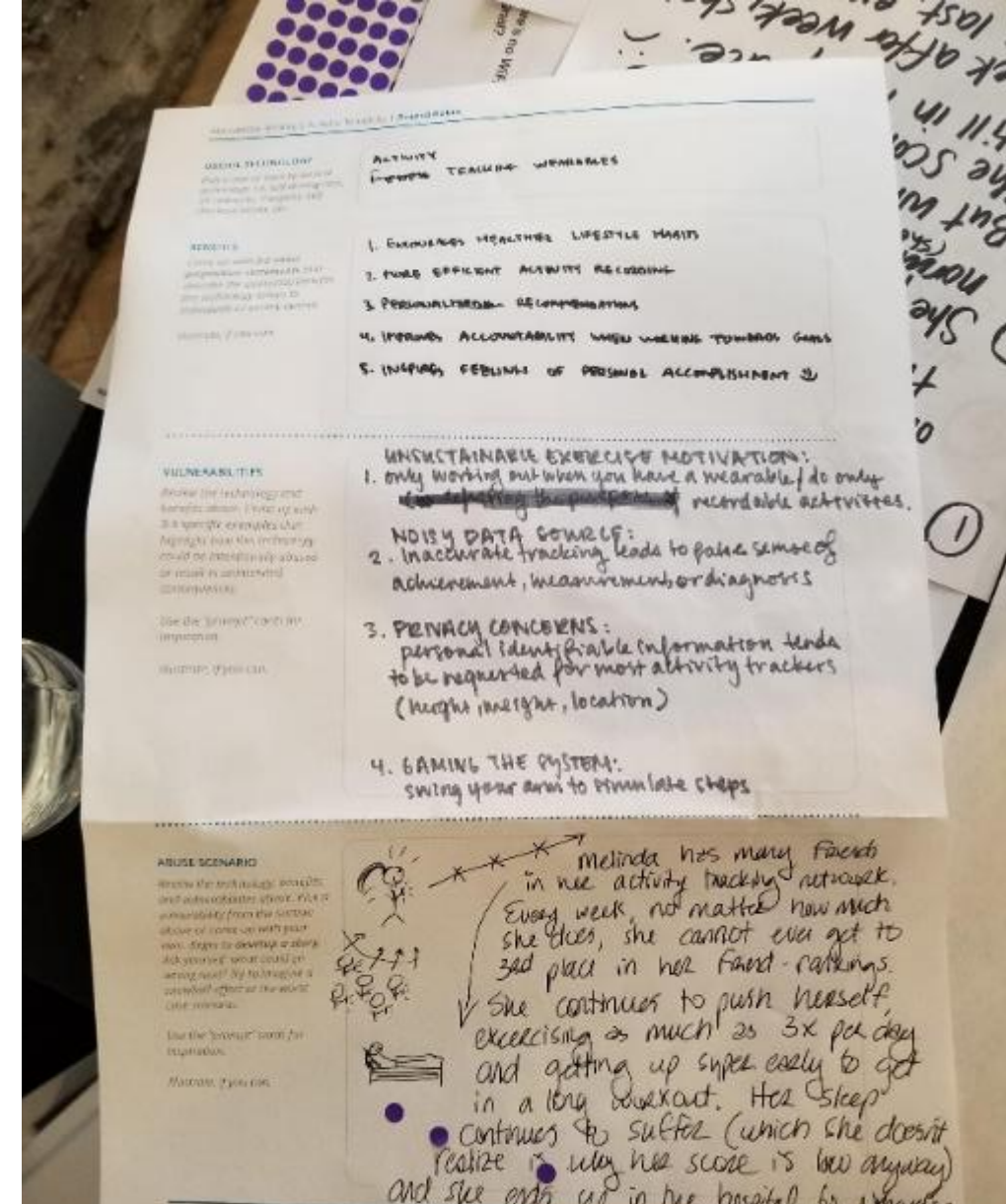
How tech could be misused

3) Abuse scenario

Provocation via prompt statements

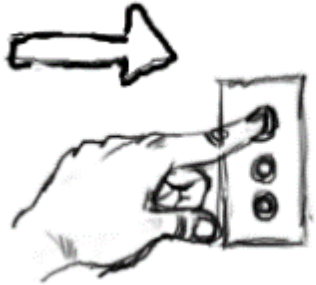
UX in the Age of Abusability. The role of Composition, Collaboration, and Craft in building ethical products. Dan Brown. Sep 18, 2018. <https://greenonions.com/ux-in-the-age-of-abusability-797cd01f6b13>

Photo from workshop organized by Anna Abovyan, Theora Kvitka and Allison Cosby of the Pittsburgh IxDA Chapter for World Interaction Design Day 2019.

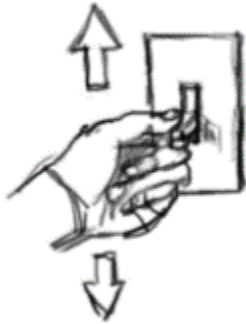


Template by: Anna Abovyan & Allison Cosby, IxDA Pittsburgh, Sep 2019

Prototype: Make Informed Design Choices



Button - Push



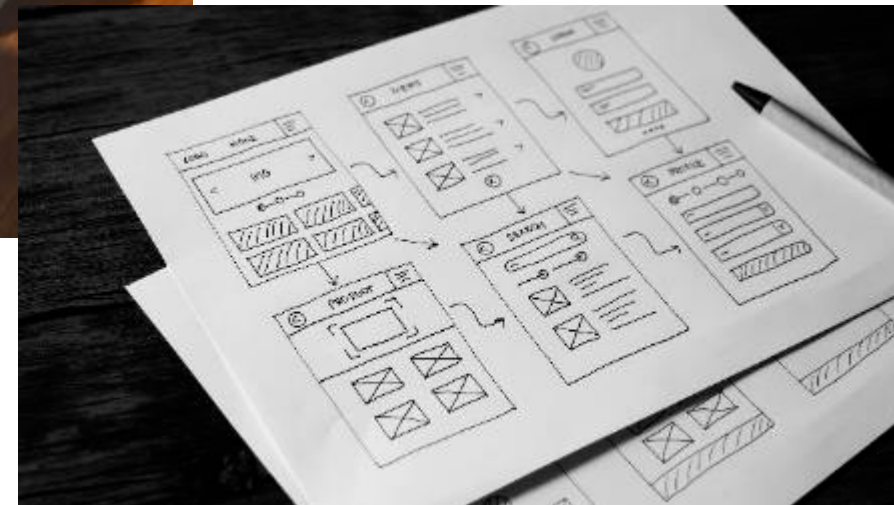
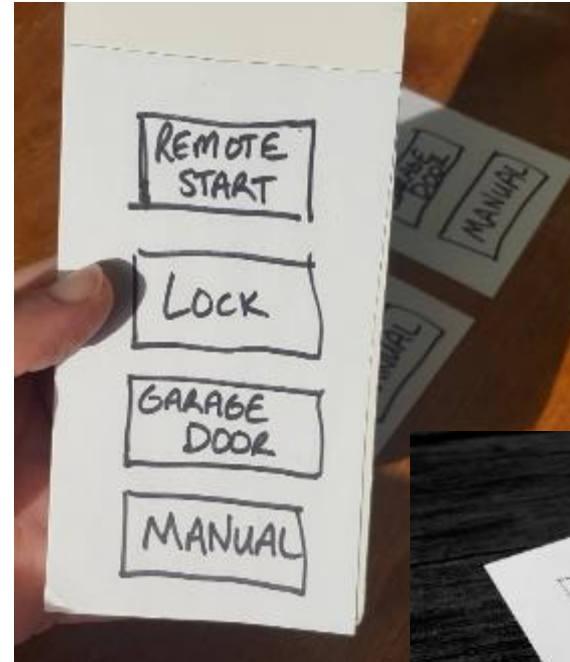
Switch - Flip



Knob - Rotate



Light Feedback



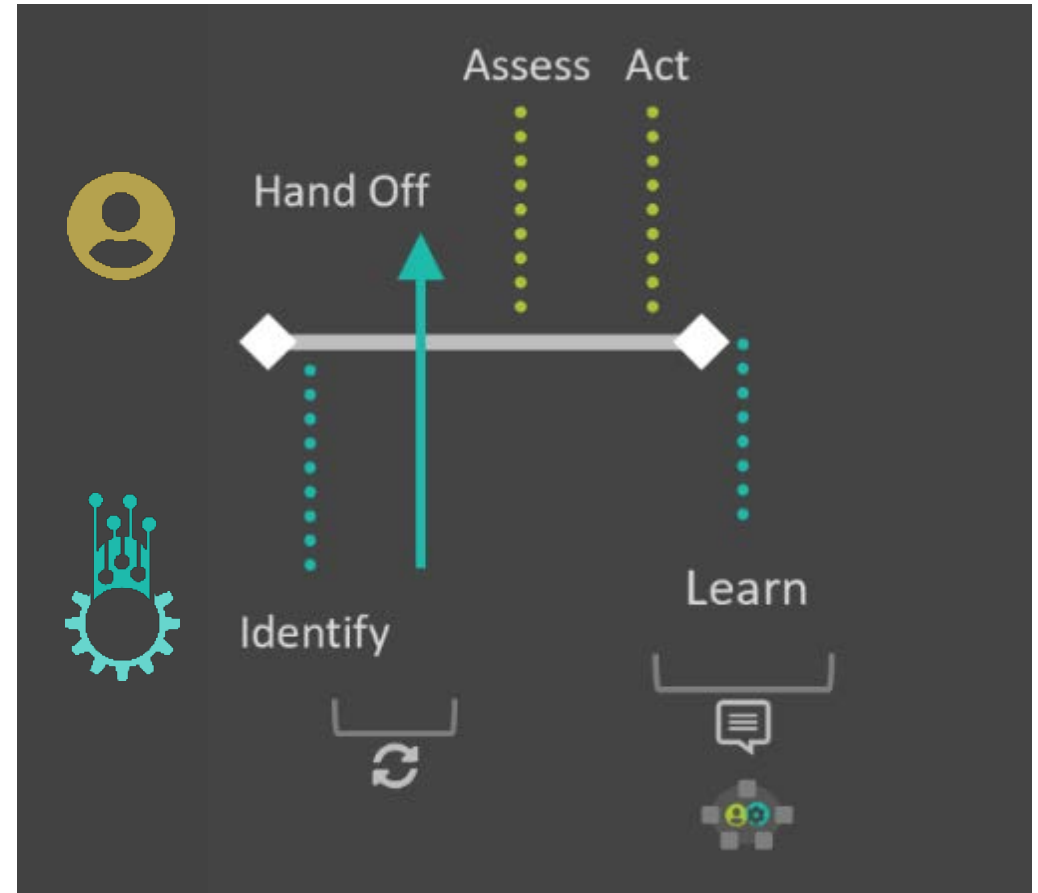
Drawings of Affordance: <http://paaralan.blogspot.com/2010/09/affordance-and-educational-games.html>

Significant Decisions

Made by system

- explained
- able to be overridden
- appealable and reversible

Responsibilities are explicitly defined between people and systems.



Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020.

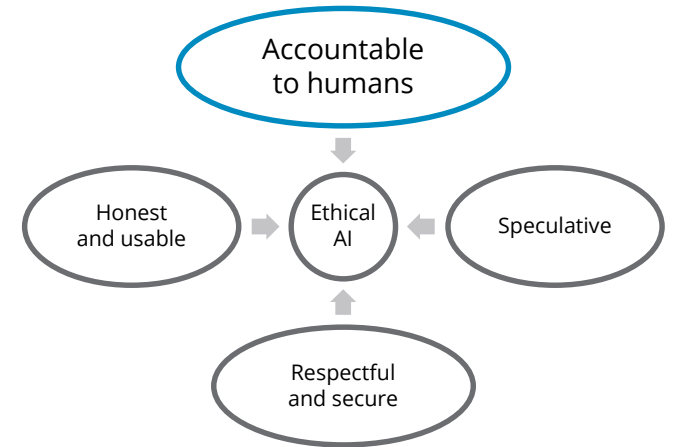
How IAs Can Shape the Future of Human-AI Collaboration. Carol Smith and Duane Degler. Presented on April 28-30, 2021 at IAC21.

Humans are Accountable

Ensure humans have ultimate control.
Able to monitor and control risk.

A person is always responsible for final decisions:

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



Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020.

“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

Iterative Cycles: Feedback and Improvement

Test Prototype with Users



Analyze & Prioritize

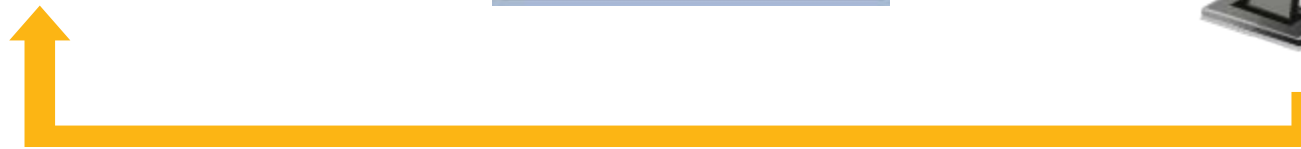
- 1 High
- 2 Medium
- 3 Low



Iterate



Repeat



Reward team members for finding ethics bugs

**Ayanna
Howard**



Let go of Some of the Numbers

Rely More on Qualitative Information





TOMATO
Solanum lycopersicum
AVG. 123 grams - 22 kcal
Nutrition Facts: Tomatoes, and tomato products - 100 grams
Calories 18
Water 95 %
Protein 0.9 g
Carbs 3.4 g
Sugar 2.6 g
Fiber 1.2 g
Pot 0.2 g
Saturated 0.03 g
Monounsaturated 0.03 g
Polyunsaturated 0.02 g
Omega-3 0 g
Omega-6 0.06 g

What do the people who will use the system expect?

Provide Evidence



Design AI to work with, and for, people

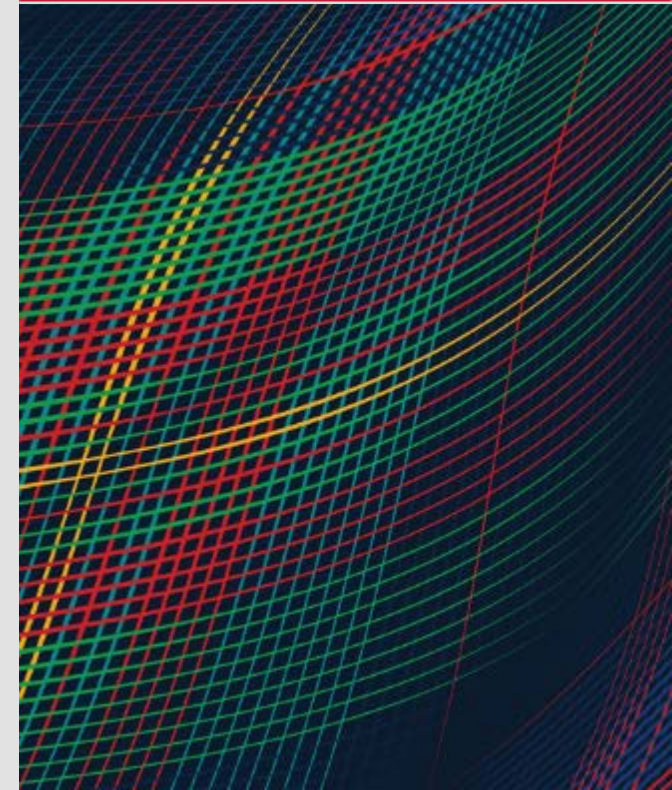


Carol J. Smith

AI Division Trust Lab Lead
Principal Research Scientist

Email: cjsmith@andrew.cmu.edu

LinkedIn: <https://www.linkedin.com/in/caroljsmith/>



Appendix



Adopt Technology Ethics

Harmonize cultural variations.

Balance to pace of change.

Explicit permission to consider and question breadth of implications.



An initiative of Université de Montréal



Prompt conversations

Checklists, frameworks, and guidelines – pair with technical ethics.

- Bridge gaps between “do no harm” and reality
- 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 affected people
- ☐ mitigation plans for manage the identified speculative risk

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

Team Signatures and Date

About the SEI
The Software Engineering Institute is a federally funded research and development center (STRO) that works with defense and government agencies, industry and academia to advance the state of the art in software engineering. Systems are built to health the public interest. Federal Government contracts by the SEI USA Federal Acquisition Regulation (FAR) are managed, performed, software acquisition, and software development.

©2013 Carnegie Mellon University. 0271 - © 10/7/2019 | S 12-092319

DEFENSE INNOVATION UNIT



RESPONSIBLE AI GUIDELINES IN PRACTICE

LESSONS LEARNED FROM THE DIU AI PORTFOLIO

BY: JARED DUNNMON, BRYCE GOODMAN, PETER KIRECHU, CAROL SMITH, & ALEXANDREA VAN DEUSEN
ACCELERATING COMMERCIAL TECHNOLOGY FOR NATIONAL SECURITY

Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog, March 9, 2020. Checklist and Agreement - Downloadable PDF: <https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620>
Defense Innovation Unit. Artificial Intelligence Portfolio, Responsible AI Guidelines. <https://www.diu.mil/responsible-ai-guidelines>

Tools to Support Conversations for Understanding

Pair DoD Ethical Principles for AI (or another set) with frameworks and tools that provoke discussion on relevant topics.



Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020. Checklist and Agreement - Downloadable PDF: <https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620>

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

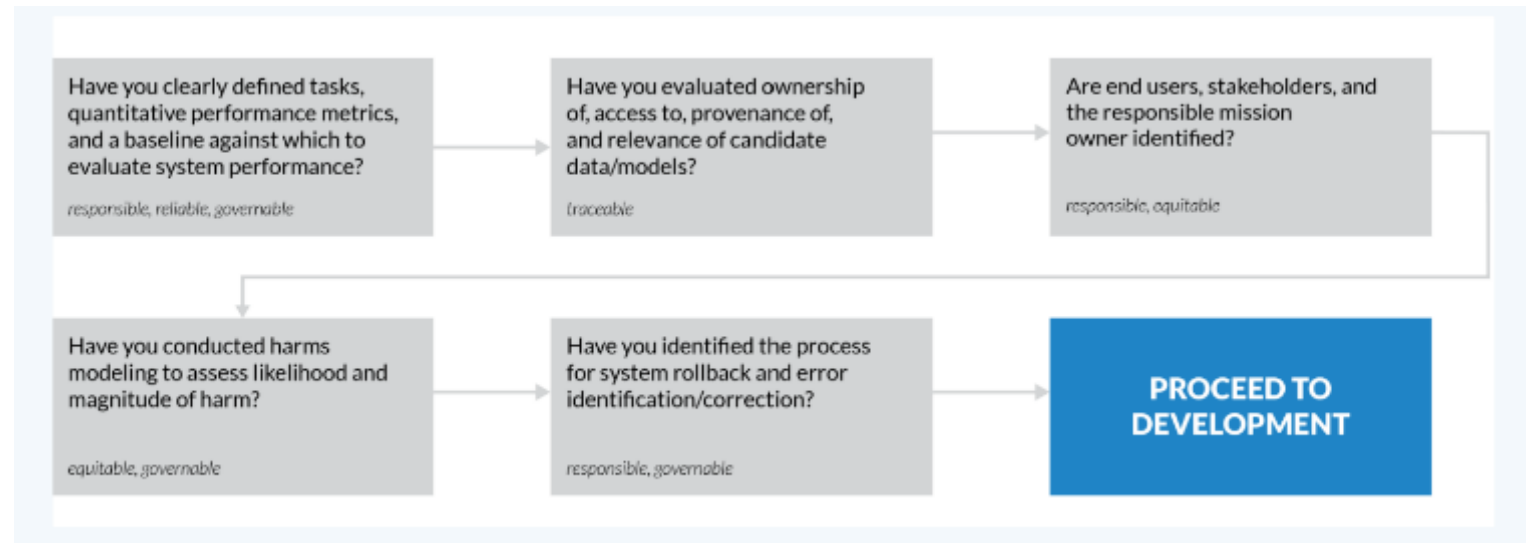
Publications



Defense Innovation Unit



RAI Report, Guidelines, Worksheets, and Workshops



Defense Innovation Unit. Artificial Intelligence Portfolio, Responsible AI Guidelines. <https://www.diu.mil/responsible-ai-guidelines>

Usable Hazard Analysis for AI Engineering

Support teams making complex systems, in early risk identification.

Exploring Opportunities in Usable Hazard Analysis Processes for AI Engineering

Nikolas Martelaro,¹ Carol J. Smith,² Tamara Zilovic¹

HCI Institute - Carnegie Mellon University,¹ Software Engineering Institute, Carnegie Mellon University²
nikmart@cmu.edu, cjsmith@sei.cmu.edu, tzilovic@andrew.cmu.edu

Abstract

Embedding artificial intelligence into systems introduces significant challenges to modern engineering practices. Hazard analysis tools and processes have not yet been adequately adapted to the new paradigm. This paper describes initial research and findings regarding current practices in AI-related hazard analysis and on the tools used to conduct this work. Our goal with this initial research is to better understand the needs of practitioners and the emerging challenges of considering hazards and risks for AI-enabled products and services. Our primary research question is: *Can we develop new structured thinking methods and systems engineering tools to support effective and engaging ways for preemptively considering failure modes in AI systems?* The preliminary findings from our review of the literature and interviews with practitioners highlight various challenges around

implications for the organizations that develop these products. While the use of new technologies always comes with the possibility of unintended consequences, we believe that many of these examples could have been prevented through strategic and thoughtful consideration when these systems are being designed and engineered.

Within systems engineering, strategies for hazard analysis can be used by teams to identify risks and potential failures with the goal of developing more robust and safe engineered systems. While many formal hazard analysis techniques exist, these activities largely center around helping teams determine potential risks and/or sources of failure *before* products have begun the development



tinyurl.com/hazards-ai-eng

Nikolas Martelaro, Carol J. Smith, and Tamara Zilovic. 2022. Exploring Opportunities in Usable Hazard Analysis Processes for AI Engineering. Presented at 2022 AAI Spring Symposium Series Workshop on AI Engineering: Creating Scalable, Human-Centered and Robust AI Systems. arXiv:2203.15628 [cs] (March 2022).

SYSTEM | autonomous vehicle v

BUILD YOUR SCENARIO

merging
into traffic

<>

USE CASE

snowing

<>

CONTEXT OF USE

novice

<>

END USER

high

<>

LEVEL OF AUTONOMY

+

WHAT COULD GO WRONG? give me a hint

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

HOW WOULD YOU RANK THE LIKELIHOOD OF THAT EVENT?

TRIVIAL

UNLIKELY

LIKELY

Tools to Support Conversations for Understanding

Pair DoD Ethical Principles for AI (or another set) with frameworks and tools that provoke discussion on relevant topics.

Carnegie Mellon University
Software Engineering Institute

Designing Ethical AI Experiences: Checklist and Agreement

USE THIS DOCUMENT TO GUIDE THE DEVELOPMENT of accountable, developer, respect AI, secure, fit for use, 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://www.org/ai/1910/03/15>.

<p>We will design our AI system with the following in mind:</p> <ul style="list-style-type: none"> <input type="checkbox"/> Designated humans share 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"> - explained - able to be overridden - appealable and reversible 	<p>We work to specifically 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 rigorous and exhaustive research unexcused 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 affected people <input type="checkbox"/> mitigation plans for managing the identified consequences <p>We value respect and equality:</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 engineering trust:</p> <ul style="list-style-type: none"> <input type="checkbox"/> The purpose, functions, and limits of the AI system are explained in plain language <input type="checkbox"/> Data sources have a clear path, documented sources, and biases are known and explicitly stated <input type="checkbox"/> Algorithms and models are appropriate and verifiable <input type="checkbox"/> Confidence and consent are informed for humans to take decisions on <input type="checkbox"/> Transparent justification for recommendations and decisions is provided <input type="checkbox"/> Design forecasts and traceable 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 (SEI) is a not-for-profit organization that provides research, education, and training in software engineering. SEI is a part of Carnegie Mellon University. For more information, visit www.sei.cmu.edu.

Contact Us
1515 Lehigh Avenue, Pittsburgh, PA 15261-0001
412-263-1000
se@sei.cmu.edu
www.sei.cmu.edu

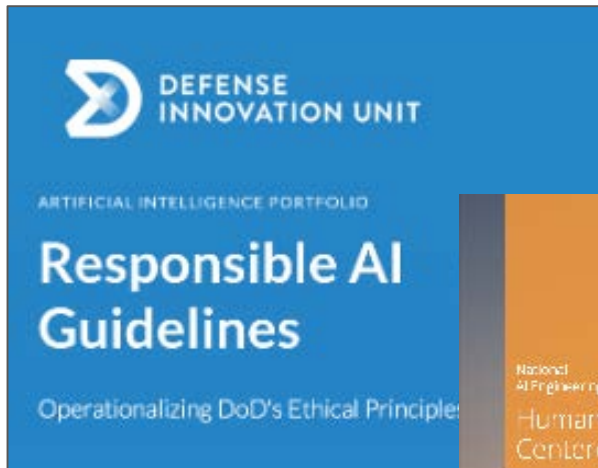
© 2020 Carnegie Mellon University. SEI-20-0001 (03/2020)



Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020.
Checklist and Agreement - Downloadable PDF: <https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620>

Photo by Pam Sharpe https://unsplash.com/@msggrace?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

Additional Publications



J. Dunnmon, B. Goodman, P. Kirechu, C. Smith, A. Van Deusen. "Responsible AI Guidelines in Practice: Lessons Learned from the DIU AI Portfolio." DIU.

H. Barmer; R. Dzombak; M. Gaston; V. Palat; F. Redner; C. Smith; et al. (2021): "Human-Centered AI." SEI, CMU.

- Blog: [Contextualizing End-User Needs: How to Measure the Trustworthiness of an AI System](#)
- Checklist: [Designing Ethical AI Experiences: Checklist and Agreement](#)
- Whitepaper: [SEI: Human-Centered AI](#)
- Blog: [What is explainable AI?](#)
- Video: [Collaboration Conversation: Human-Centered AI](#)
- Video: [Implementing the DoD's Ethical AI Principles](#)
- Video: [Bias in AI: Impact, Challenges, and Opportunities](#)