Carnegie Mellon University
Software Engineering Institute

Letting Go of the Numbers: Measuring AI Trustworthiness

ICPRAM 2024

Carol J. Smith

Trust Lab Lead & Principal Research Scientist, AI Division

Copyright Statement

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

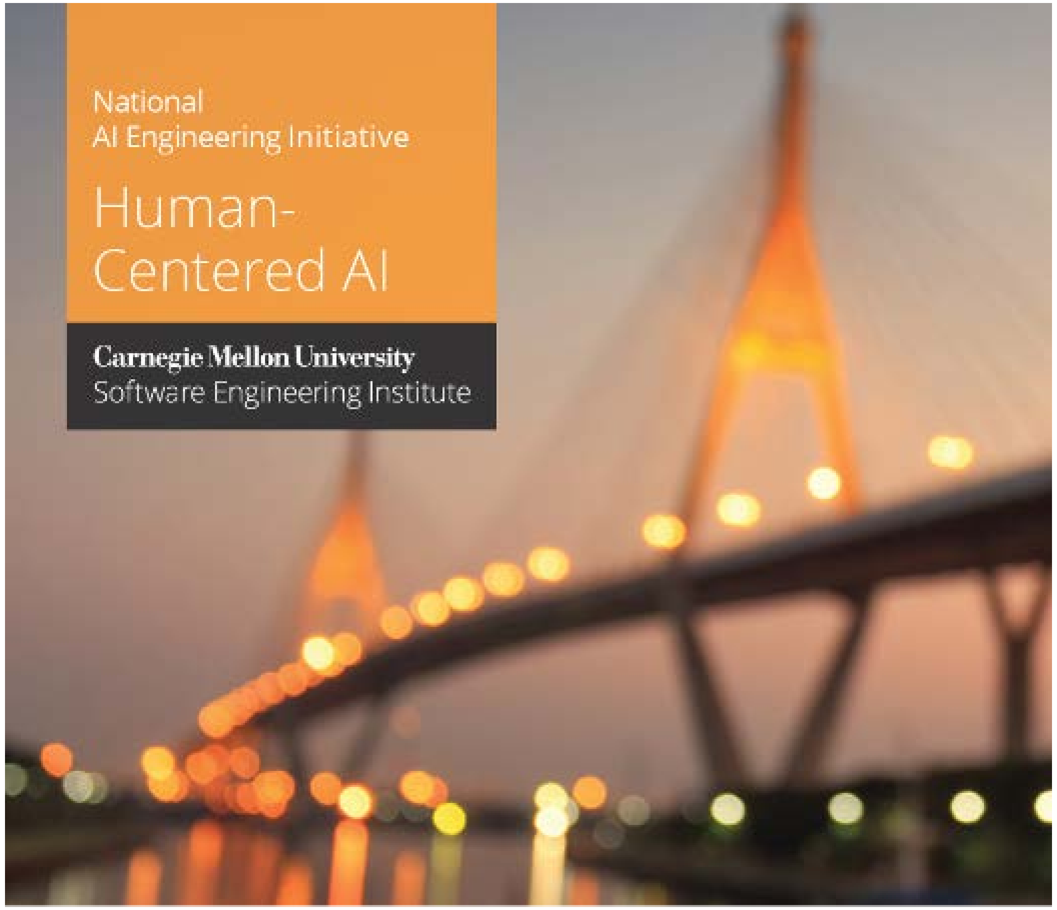
Video: Tesla Autopilot in Heavy LA Traffic by Scott Kubo https://youtu.be/m3-QzTFxoUg?t=14

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?

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

Carnegie Mellon University Software Engineering Institute

Can Anyone?

Carnegie Mellon University Software Engineering Institute

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.

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

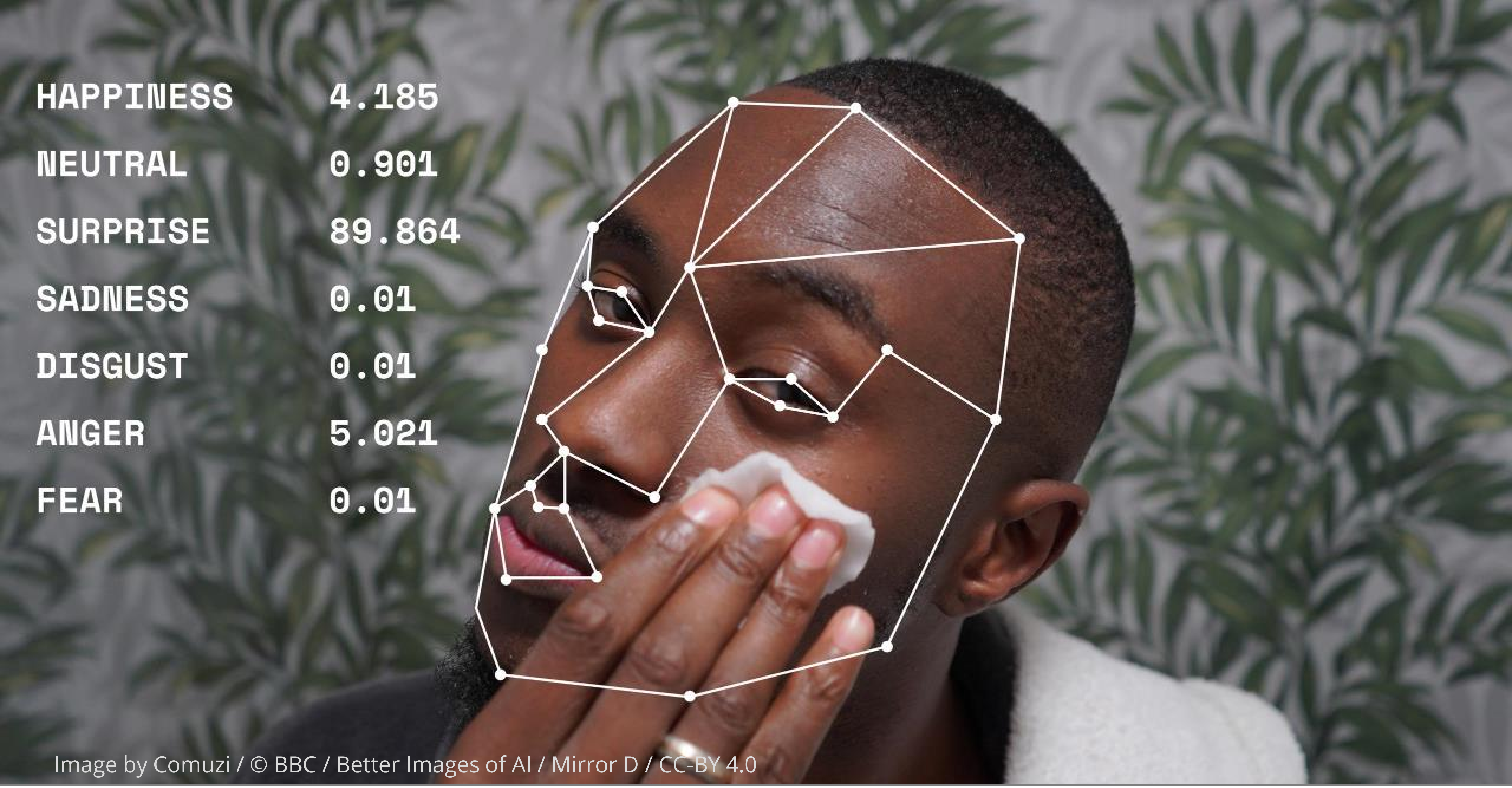
SHARE

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

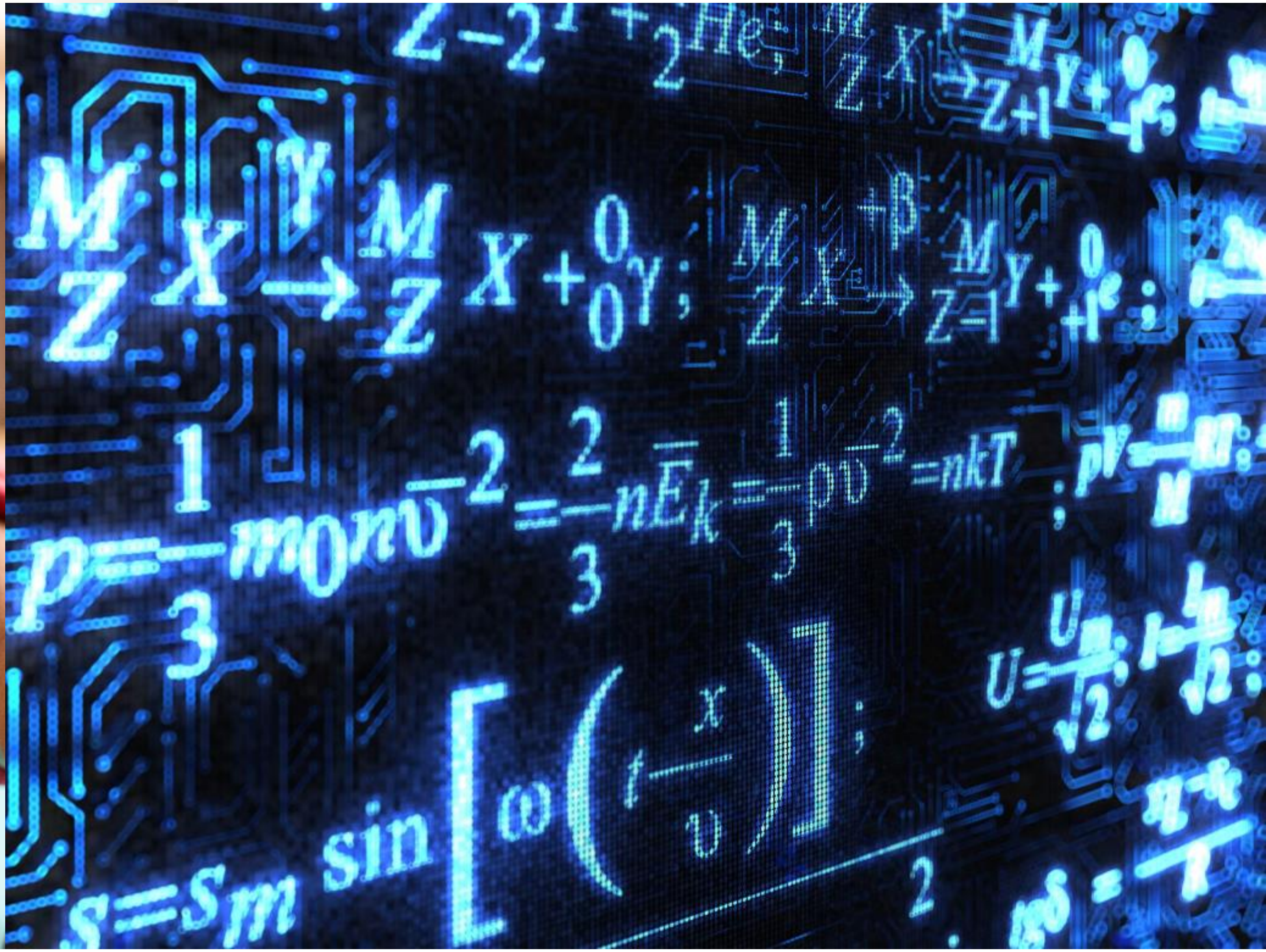
Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech. By James Vincent Jan 12, 2018. The Verge.

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

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

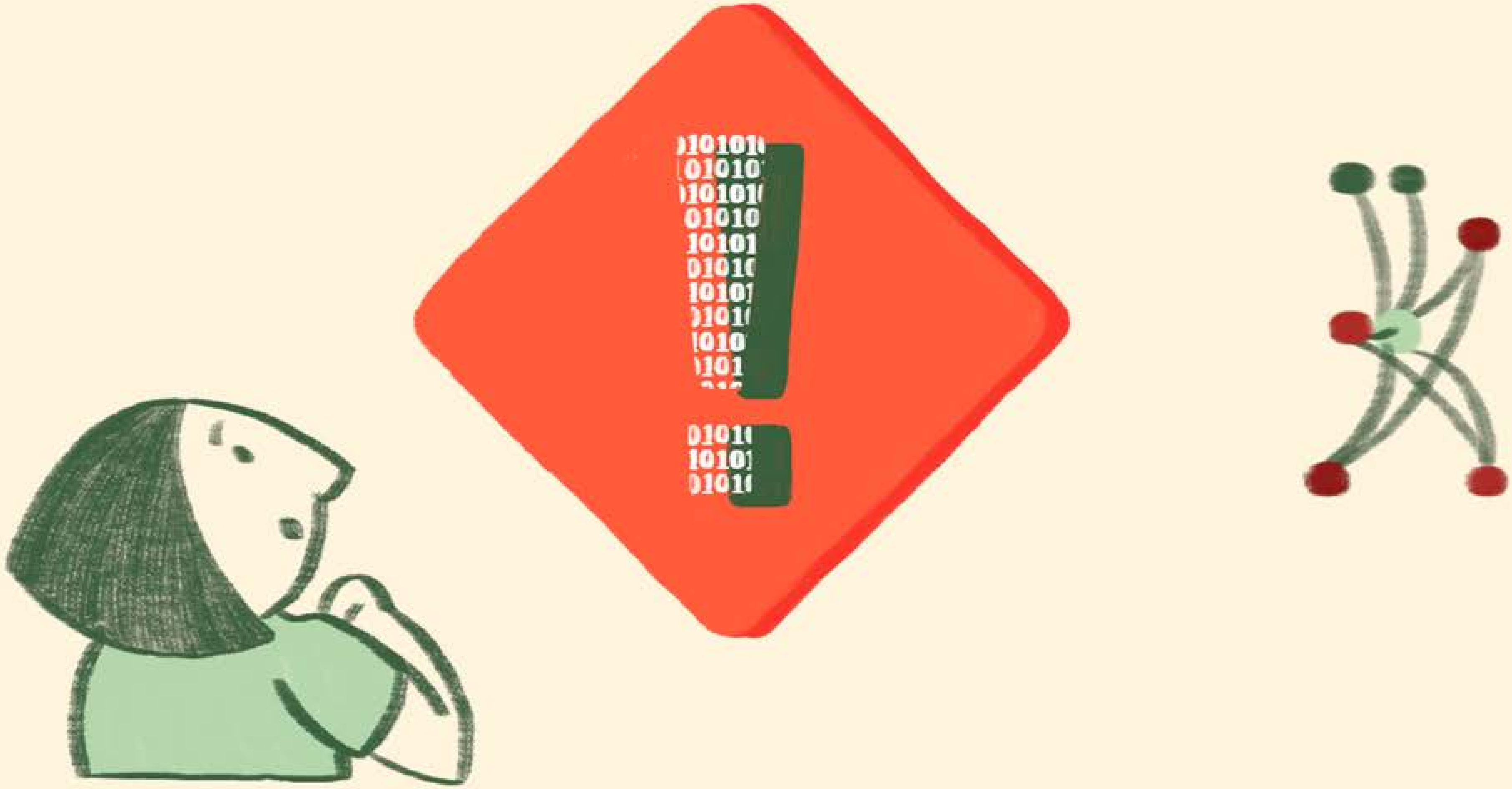
Quantitative

- 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

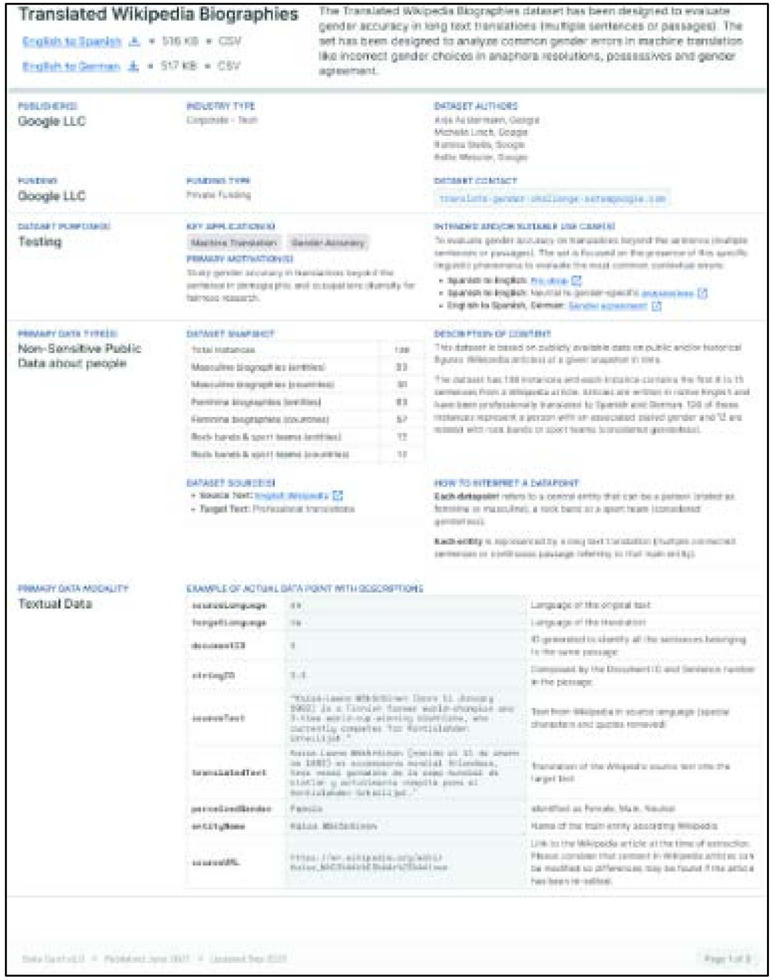
Photo: Joy Buolamwini on The Open Mind: Algorithmic Justice.
Jan 12, 2019. https://www.youtube.com/watch?v=hwHnXdoSSFY

[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.

Start with Data

Need to confirm:

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



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

Spotted Lanternfly

Carnegie Mellon University Software Engineering Institute

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

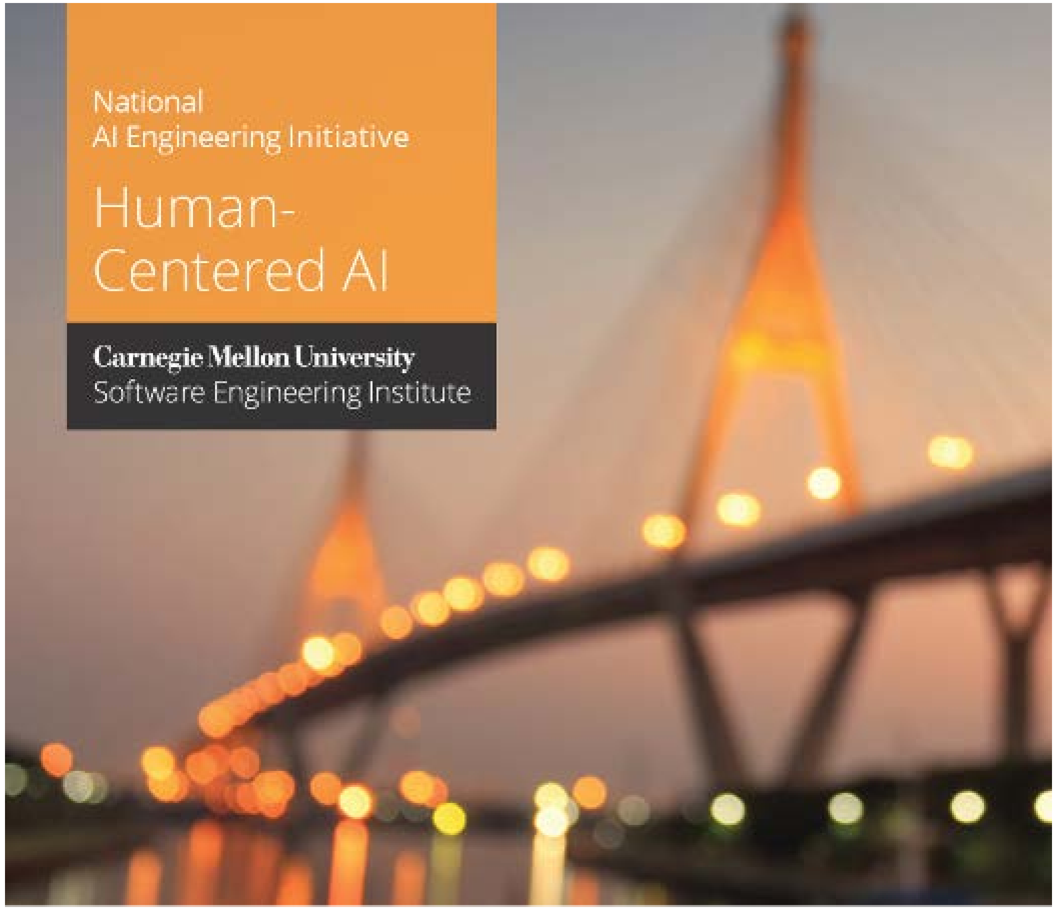
Socio-political nuance

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.

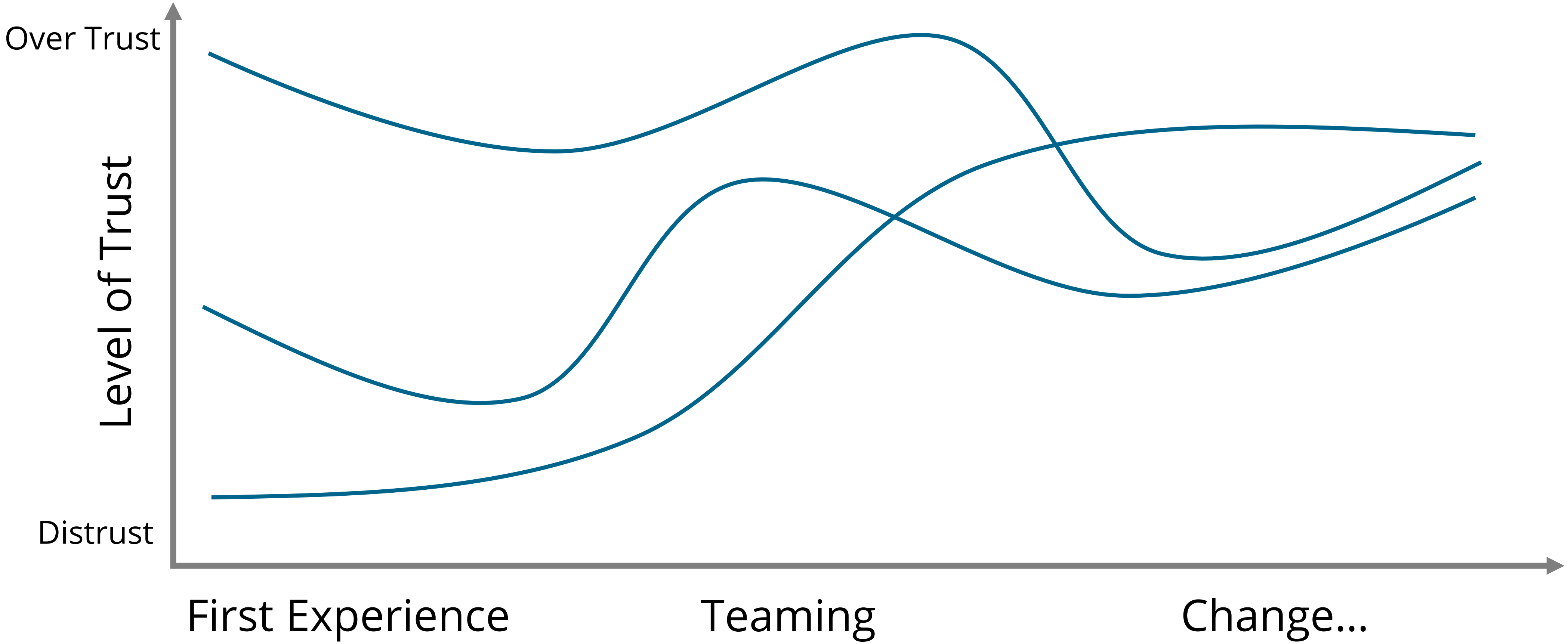
Calibrated Trust
Trust matches system capabilities - leading to appropriate use.

Over Trust
Trust exceeding system capabilities - may lead to misuse.

Rejection. ⟷ **Automation bias.**

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

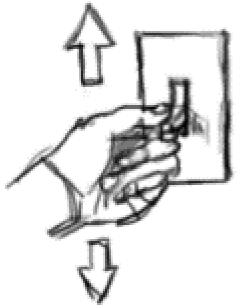
- Who will use the system?
- How well are current systems accepted?
- What are the existing issues?
- When and in what context?

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

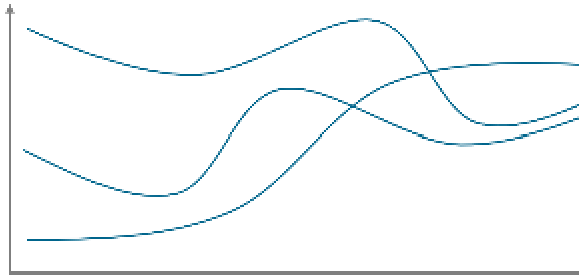


Explainability

Fairness

Likelihood of Failure

Usage



How is the AI system making decisions?

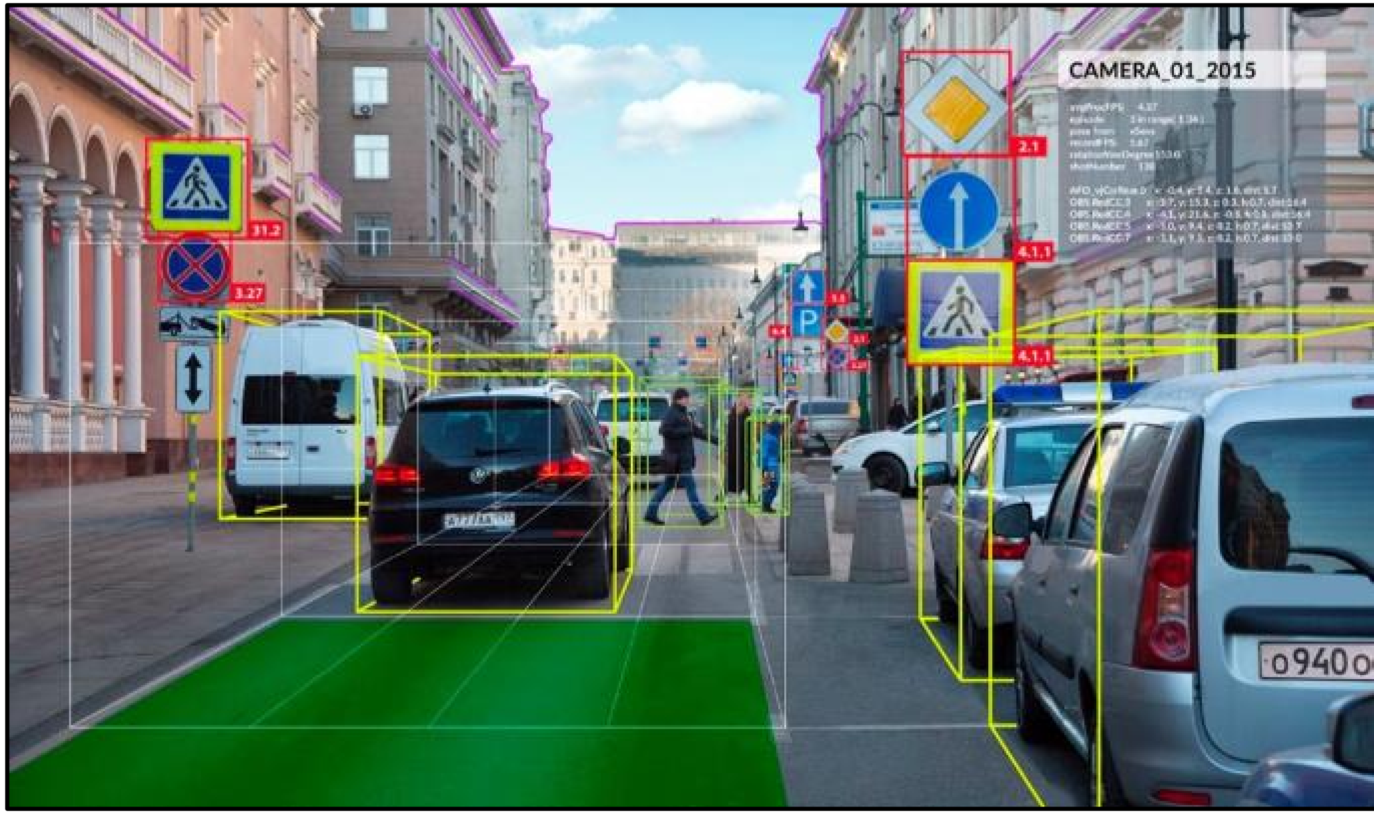
Explainability

Photo GOODFELLOW AIR FORCE BASE, TX, UNITED STATES, 11.05.2020by 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

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

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

Understanding the Likelihood of Failure

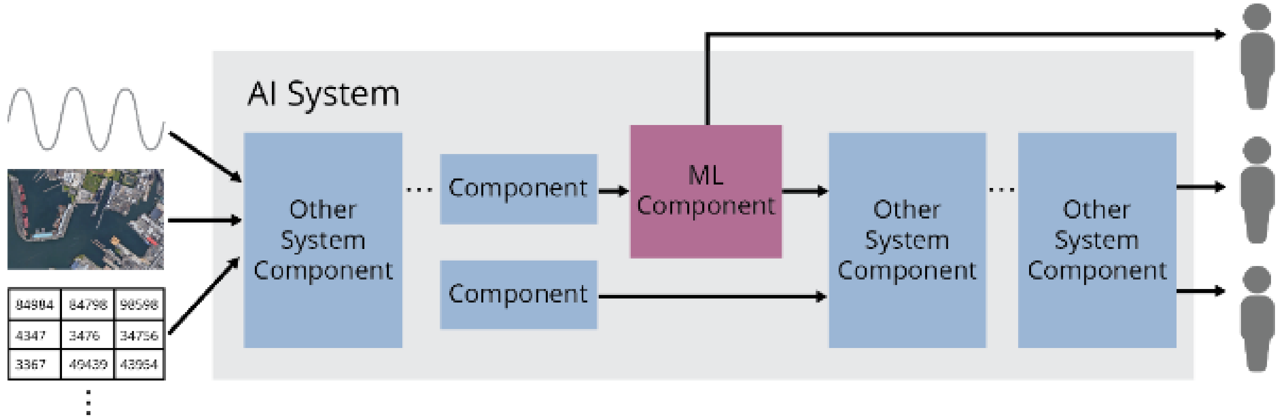
Accurate confidence measures can inform better decision making in complex contexts

People need

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

Decide what to do next

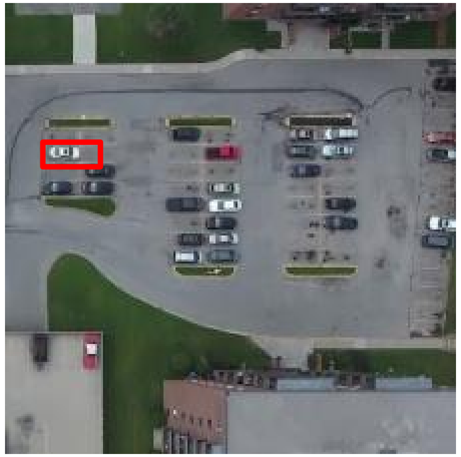
- inform other parts of the system
- alert an analyst, use another sensor, etc.



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?

Carnegie Mellon University Software Engineering Institute

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?

Carnegie Mellon University Software Engineering Institute

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

Carnegie Mellon University Software Engineering Institute

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?

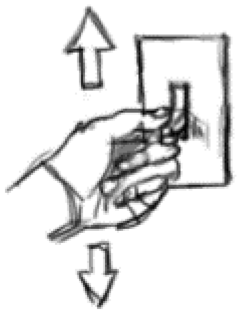
- What is the cause?
- How do we want the system to respond when it encounters new information – new situations?
- What is the appropriate way to communicate the likelihood of failure?

Testing for Trustworthiness

Carnegie Mellon University Software Engineering Institute

Measurements of Trustworthiness

Usability

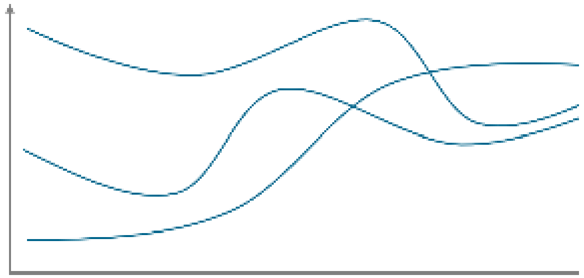


Explainability

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.

AutomotiveUI '20 Adjunct, September 21–22, 2020, Virtual Event, USA — Martelaro and Ju

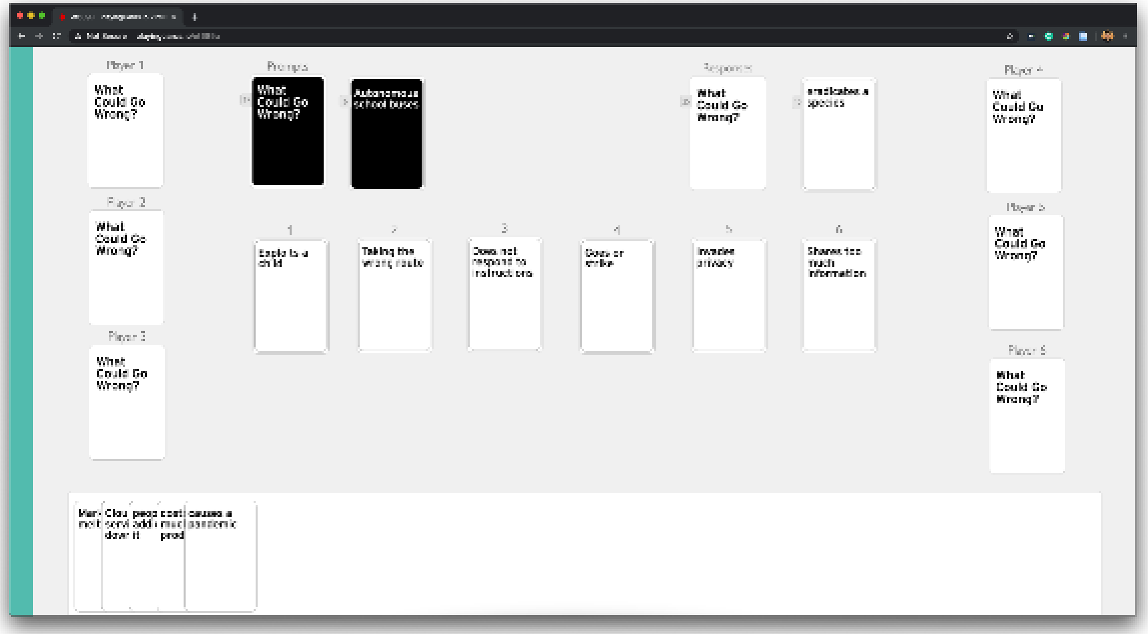


Figure 1: What Could Go Wrong? A digital card game to foster conversation around potential challenges and issues of autonomous vehicle technology.

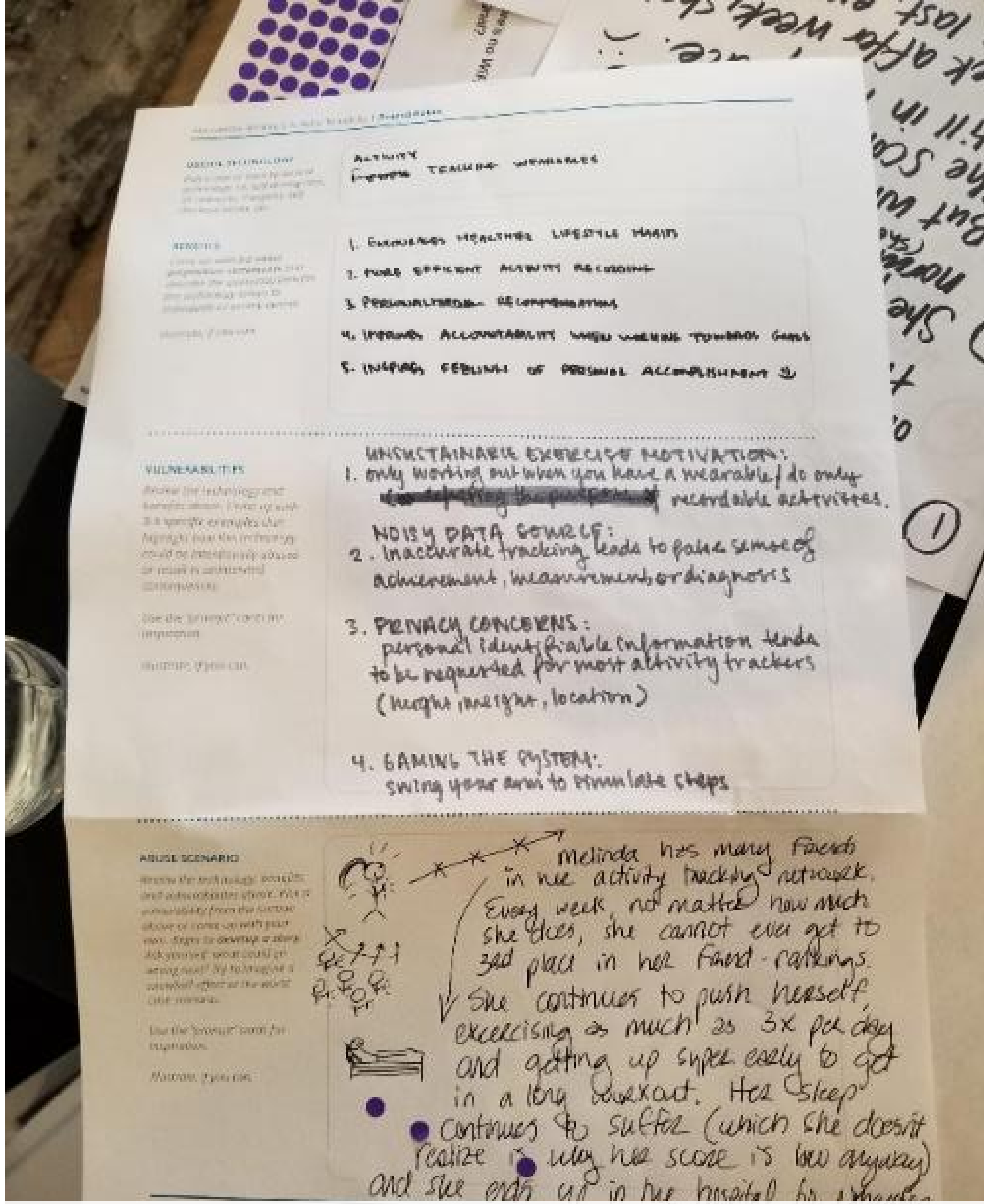
Nikolas Martelaro and Wendy Ju. 2020. What Could Go Wrong? Exploring the Downsides of Autonomous Vehicles. In 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '20). Association for Computing Machinery, New York, NY, USA, 99–101. https://doi.org/10.1145/3409251.3411734

Abusability Testing

1) **Value proposition**
   Benefits tech brings to individuals, society
2) **Vulnerabilities**
   How tech could be misused
3) **Abuse scenario**

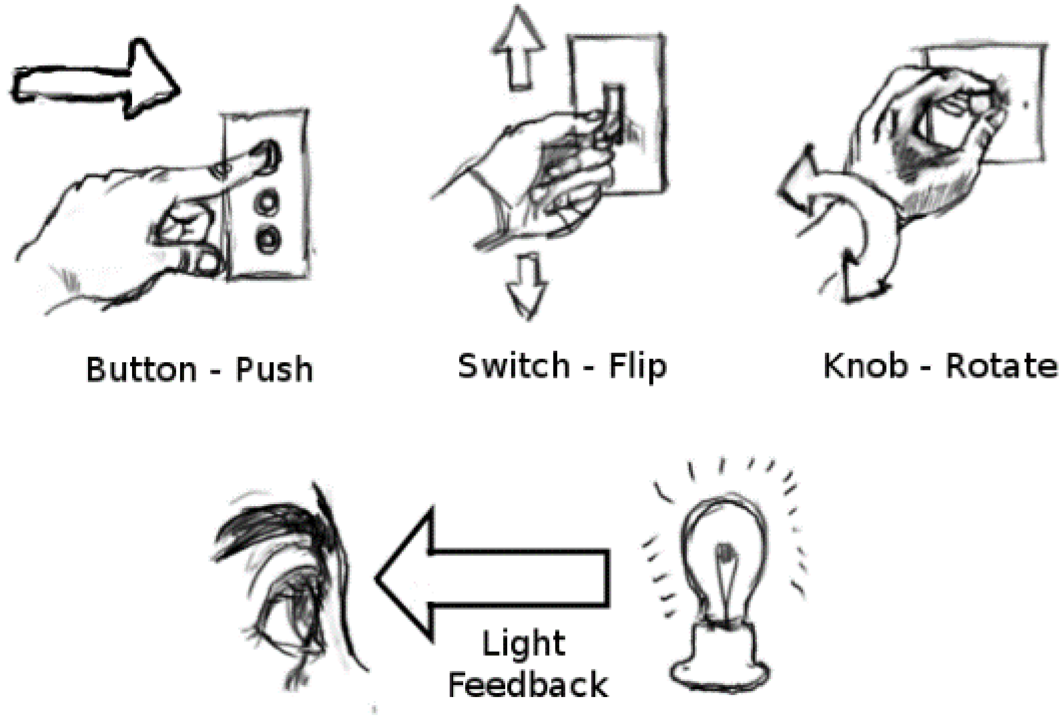
Provocation via prompt statements

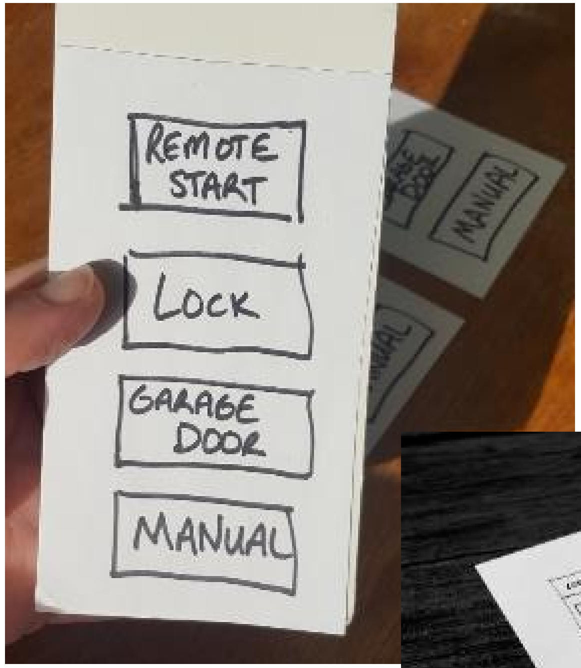
UX in the Age of Abusability. The role of Composition, Collaboration, and Craft in building ethical products. Dan Brown. Sep 18, 2018. https://greenonions.com/ux-in-the-age-of-abusability-797cd01f6b13
Photo from workshop organized by Anna Abovyan, Theora Kvitka and Allison Cosby of the Pittsburgh IxDA Chapter for World Interaction Design Day 2019.

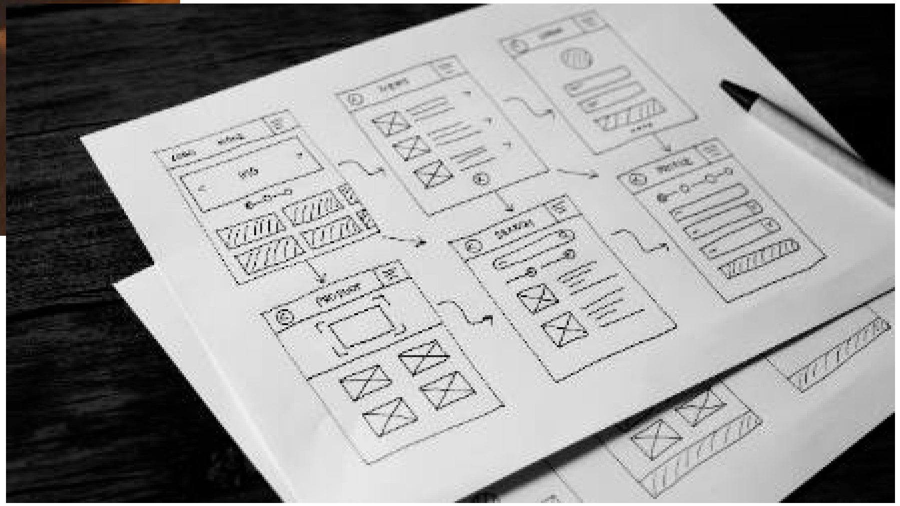


Template by: Anna Abovyan & Allison Cosby, IxDA Pittsburgh, Sep 2019

Prototype: Make Informed Design Choices





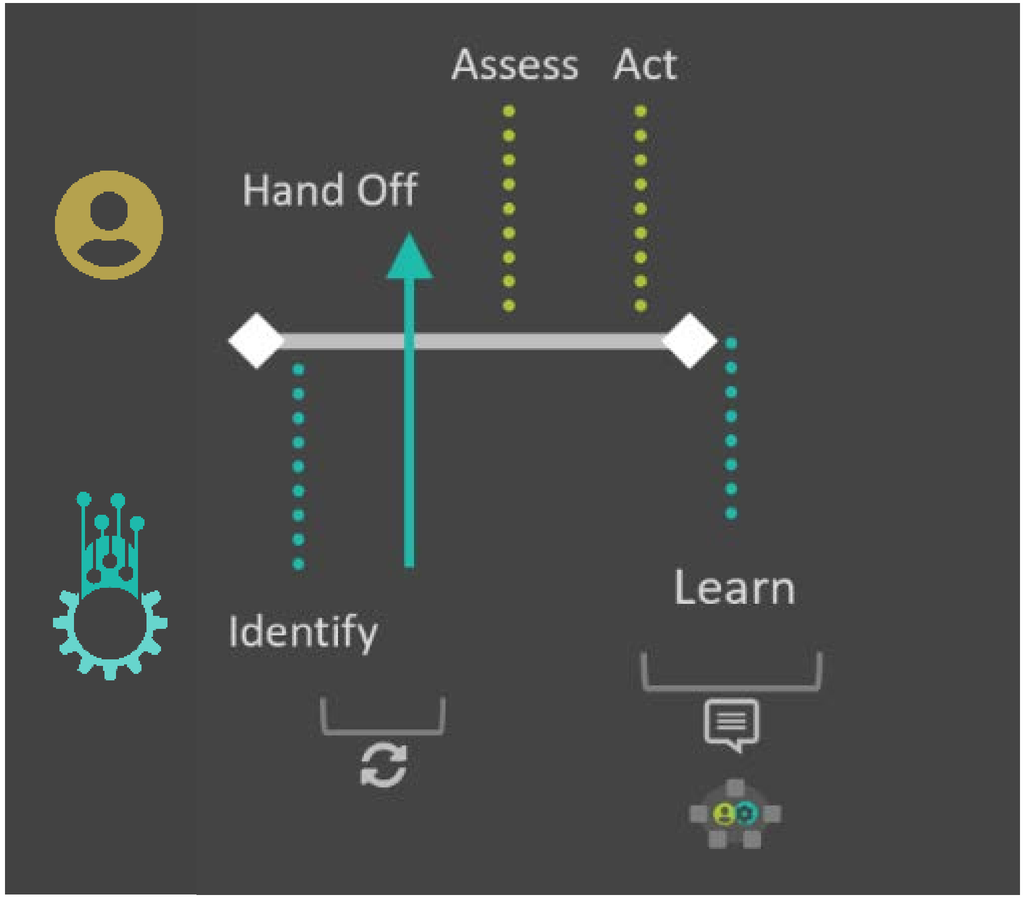


Drawings of Affordance: http://paaralan.blogspot.com/2010/09/affordance-and-educational-games.html

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.

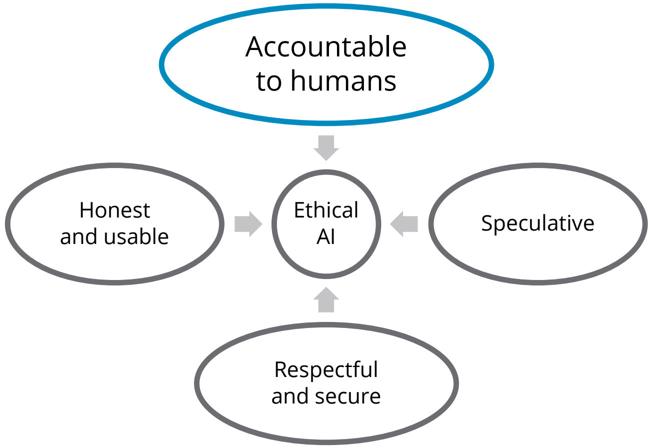
How IAs Can Shape the Future of Human-AI Collaboration. Carol Smith and Duane Degler. Presented on April 28-30, 2021 at IAC21.

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.

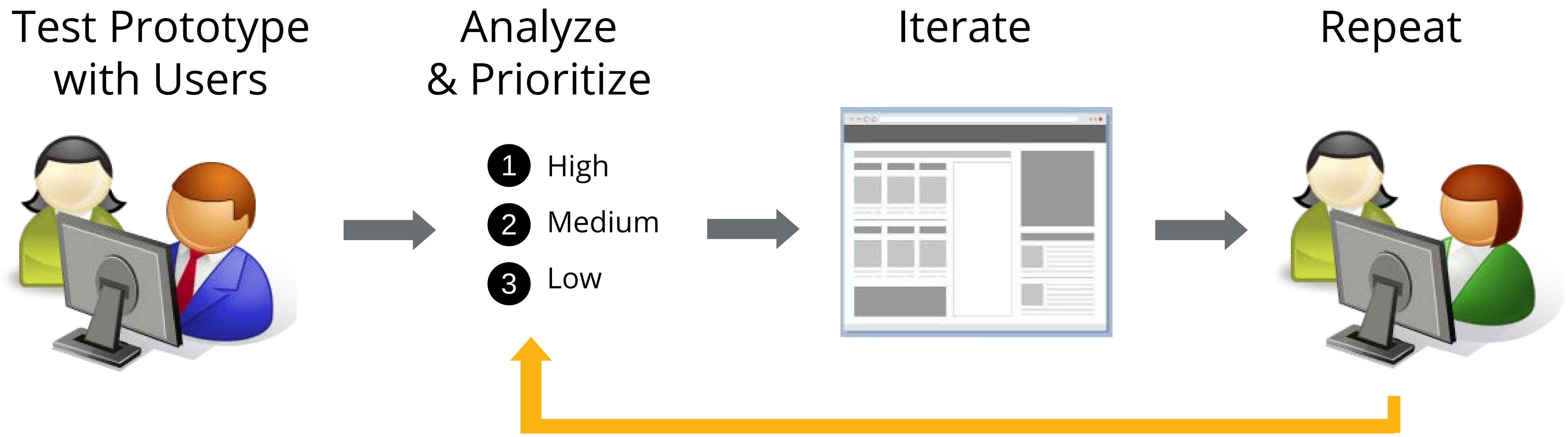
"Ensure humans can unplug the machines"

– 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



Reward team members for finding ethics bugs

Let go of Some of the Numbers

Rely More on Qualitative Information

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/

Carnegie Mellon University
Software Engineering Institute

Appendix

Carnegie Mellon University Software Engineering Institute

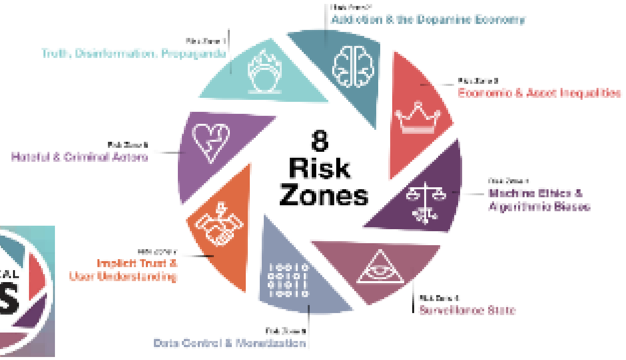
Adopt Technology Ethics

Harmonize cultural variations.

Balance to pace of change.

Explicit permission to consider and question breadth of implications.

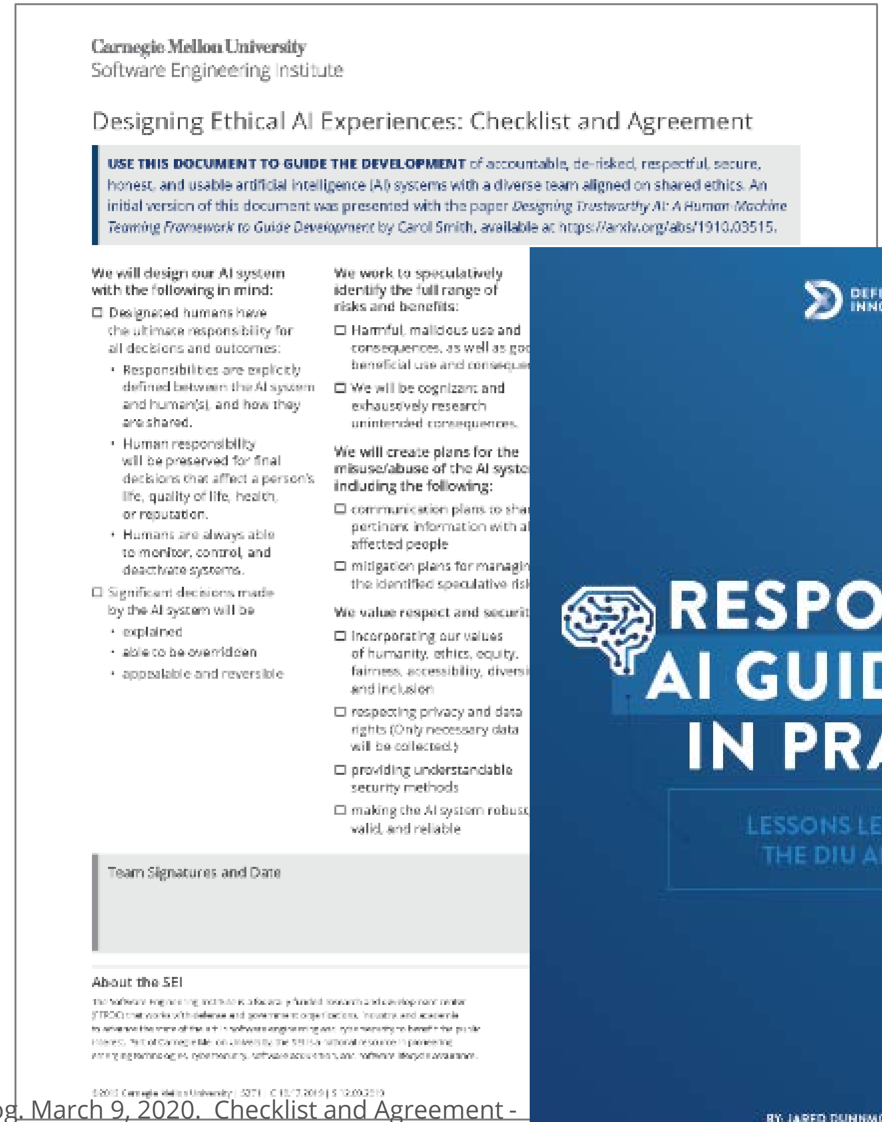
An initiative of Université de Montréal



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 goo... beneficial use and consequ...
- We will be cognizant and exhaustively research unintended consequences.

We will create plans for the misuse/abuse of the AI syste... including the following:

- communication plans to sha... pertinent information with a... affected people
- mitigation plans for managi... the identified speculative ris...

We value respect and securit...

- Incorporating our values of humanity, ethics, equity, fairness, accessibility, diversi... and inclusion
- respecting privacy and data rights (Only necessary data will be collected.)
- providing understandable security methods
- making the AI system robus... valid, and reliable

Team Signatures and Date

About the SEI

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

Carnegie Mellon University Software Engineering Institute

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

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

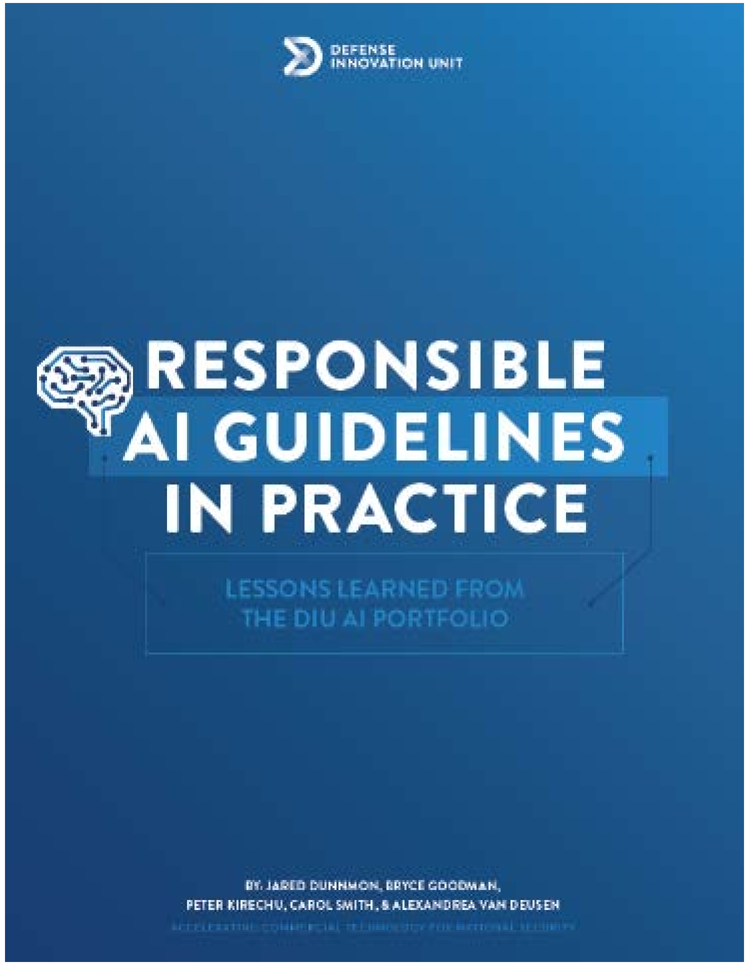
Photo by Pam Sharpe https://unsplash.com/@msgrace?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText On Unsplash https://unsplash.com/s/photos/business-woman-smiling?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText

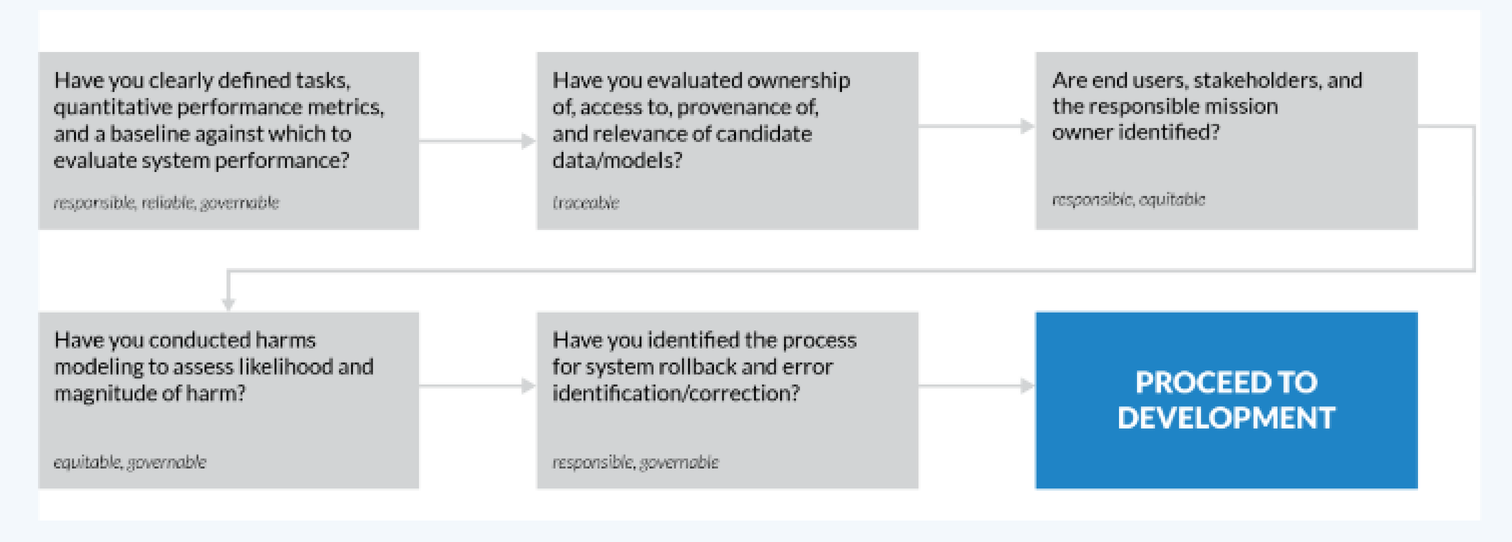
Publications

Carnegie Mellon University Software Engineering Institute

Defense Innovation Unit

RAI Report, Guidelines, Worksheets, and Workshops



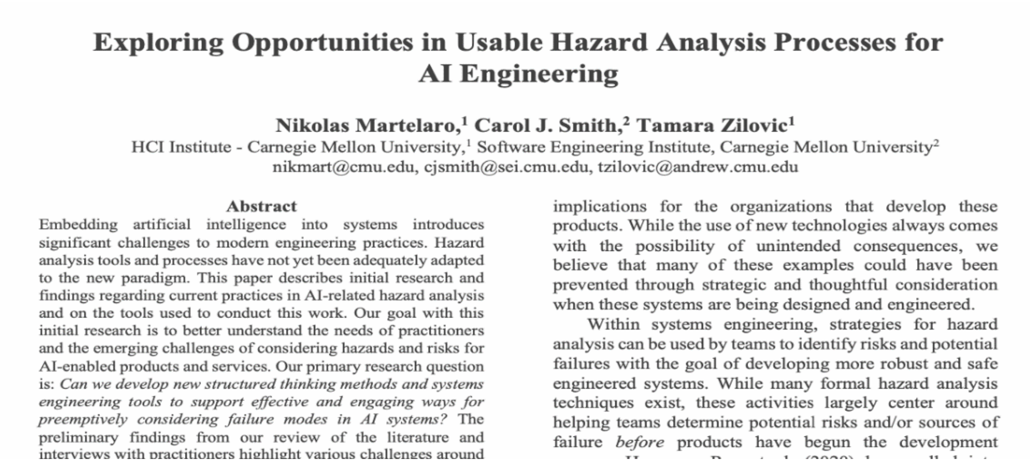


Defense Innovation Unit. Artificial Intelligence Portfolio, Responsible AI Guidelines. https://www.diu.mil/responsible-ai-guidelines

Usable Hazard Analysis for AI Engineering

Carnegie Mellon University Software Engineering Institute

Support teams making complex systems, in early risk identification.



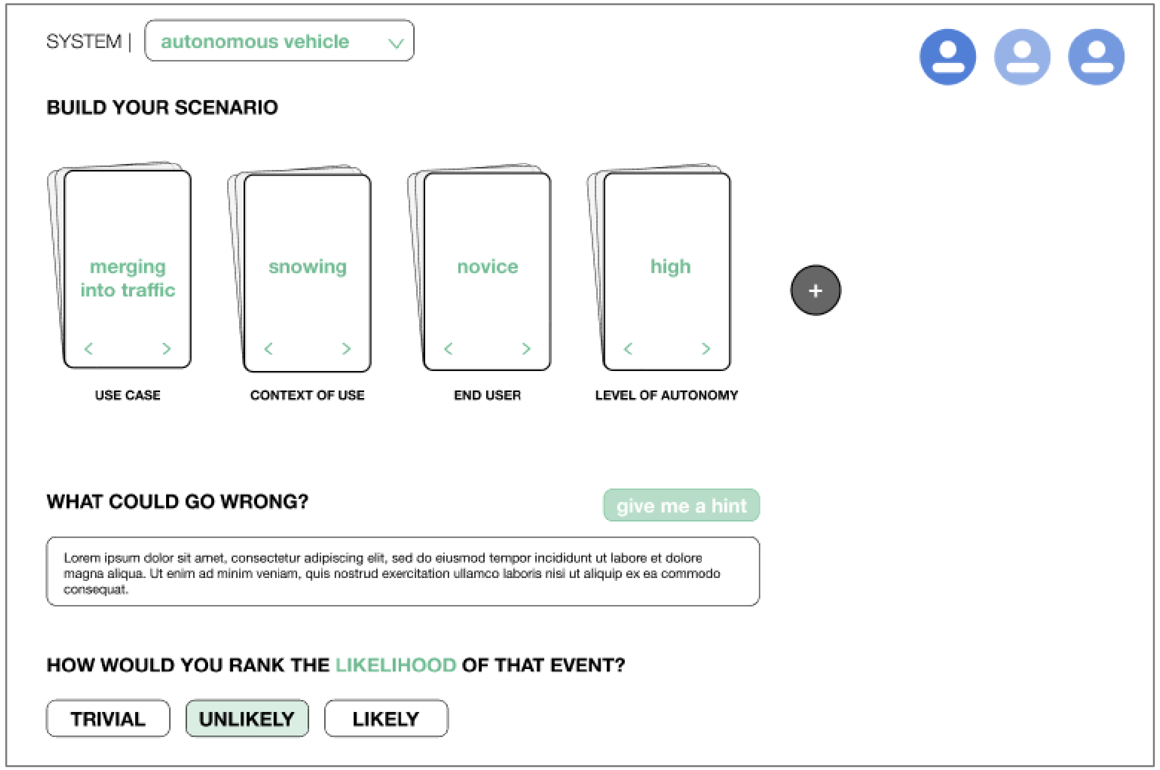
Exploring Opportunities in Usable Hazard Analysis Processes for AI Engineering

Nikolas Martelaro,[1] Carol J. Smith,[2] Tamara Zilovic[1]
HCI Institute - Carnegie Mellon University,[1] Software Engineering Institute, Carnegie Mellon University[2]
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 AAAI Spring Symposium Series Workshop on AI Engineering: Creating Scalable, Human-Centered and Robust AI Systems. arXiv:2203.15628 [cs] (March 2022).

Tools to Support Conversations for Understanding

Carnegie Mellon University
Software Engineering Institute

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

We will design our AI system with the following in mind:

We work to speculatively identify the full range of risks and benefits:

We value transparency with the goal of engendering trust:

We will create plans for the misuse/abuse of the AI system, including the following:

We value respect and security:

We value honesty and usability:

Team Signatures and Date

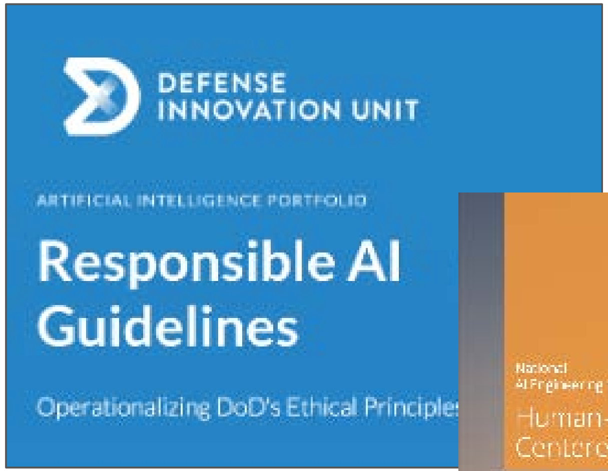
About the SEI

Contact Us

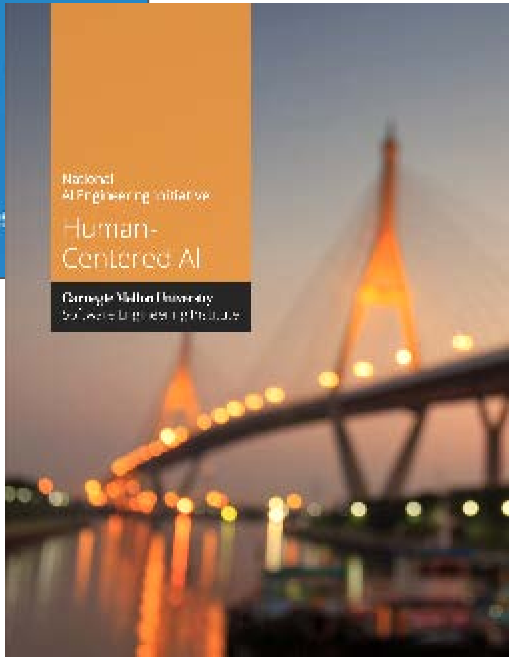
Designing Trustworthy AI for Human-Machine Teaming. By Carol Smith. Software Engineering Institute Blog. March 9, 2020.
Checklist and Agreement - Downloadable PDF: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=636620

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

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