Where to Start with AI: Ethics, Bias, and Key Considerations for Adopting AI Systems

Alex Van Deusen and Carol J. Smith
AI Division, CMU SEI
arvandeusen@sei.cmu.edu and cjsmith@sei.cmu.edu

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

The present moment is pivotal: we must act to protect our security and advance our competitiveness, seizing the initiative to lead the world in the development and adoption of transformative defense AI solutions that are ***safe, ethical, and secure****.*

–Department of Defense Artificial Intelligence Strategy 2018

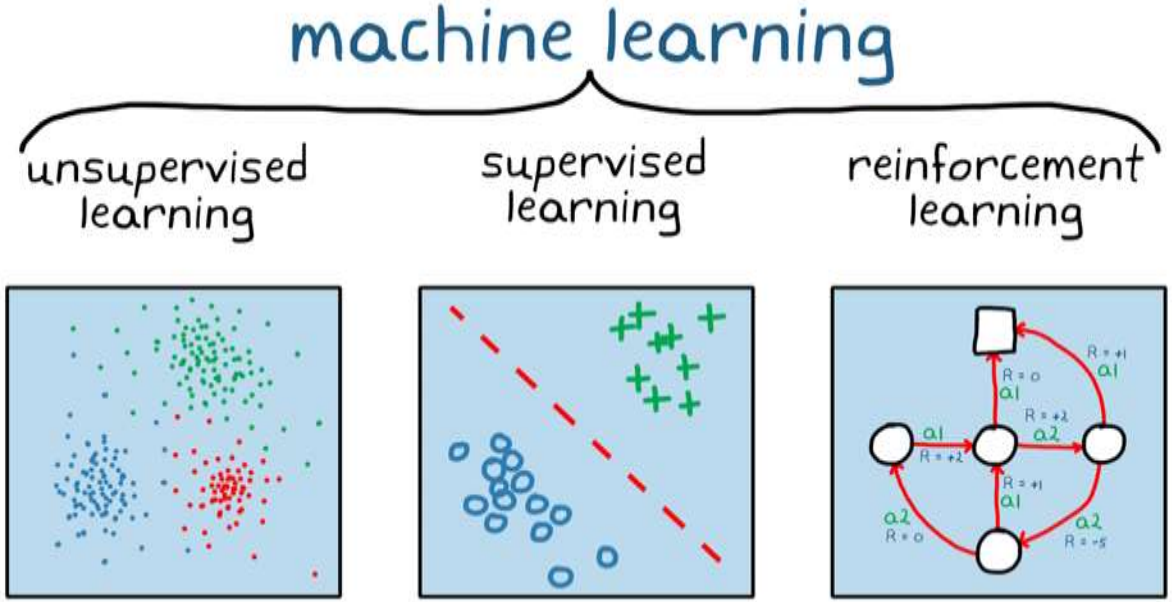
AI Primer

What is Artificial Intelligence?

AI systems can
- recognize patterns
- create predictions
- make decisions, and/or
- generate new content.

Without being
explicitly programmed to do so.

AI is typically created with machine learning (ML) methods



+ deep learning, neural networks, etc.

Image What Is Reinforcement Learning? 3 things you need to know. © 1994-2022 The MathWorks, Inc.
https://www.mathworks.com/discovery/reinforcement-learning.html

Machine Learning

Requirements for ML

1. **Data**: pre-existing, machine readable, relevant (amount vary)
2. **Math**: appropriate for data and context (statistics, probability, calculus…)
3. **Programming**: Python, C/C++, R, Java, JavaScript…

Math + programming = algorithm

Data + algorithm = ML model*

*The term model is used inconsistently. Model sometimes refers to an algorithm without data.

How can teams harness the **power** of AI systems and design them to be **valuable** to humans?

Where to Start with AI

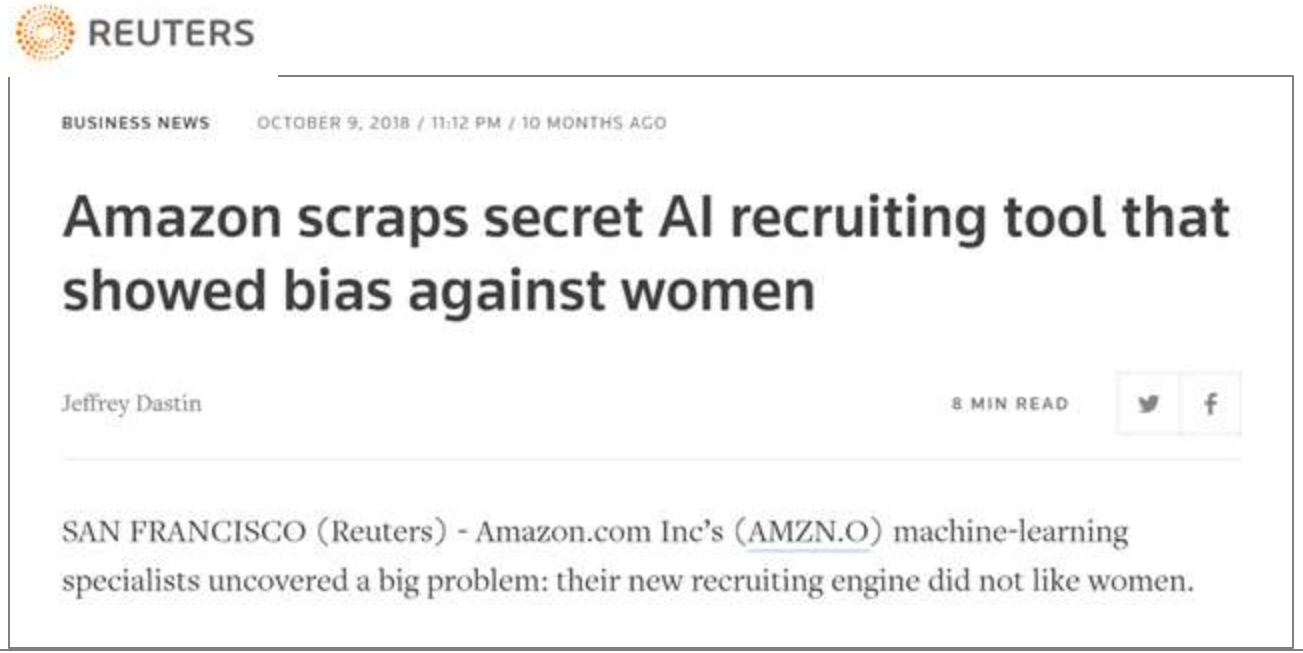
Create value with a focus on user needs

Responsible, Intentional Design

Just because you *can* doesn't mean that you *should*.

Artificial intelligence isn't always the right answer

Amazon assumed that math would reduce bias in applicant vetting.



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.

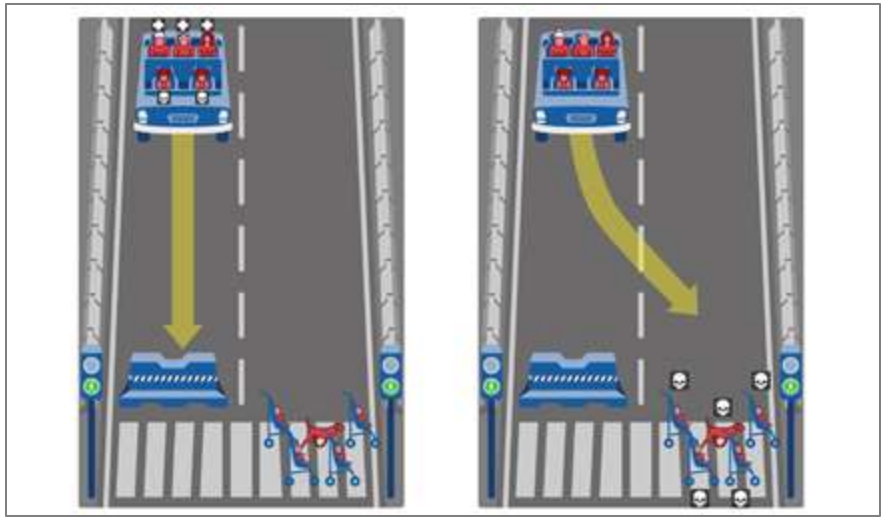
Where to Start with AI

Ethics and Bias

Technical ethics, not academic Trolley Problems

MORAL MACHINE

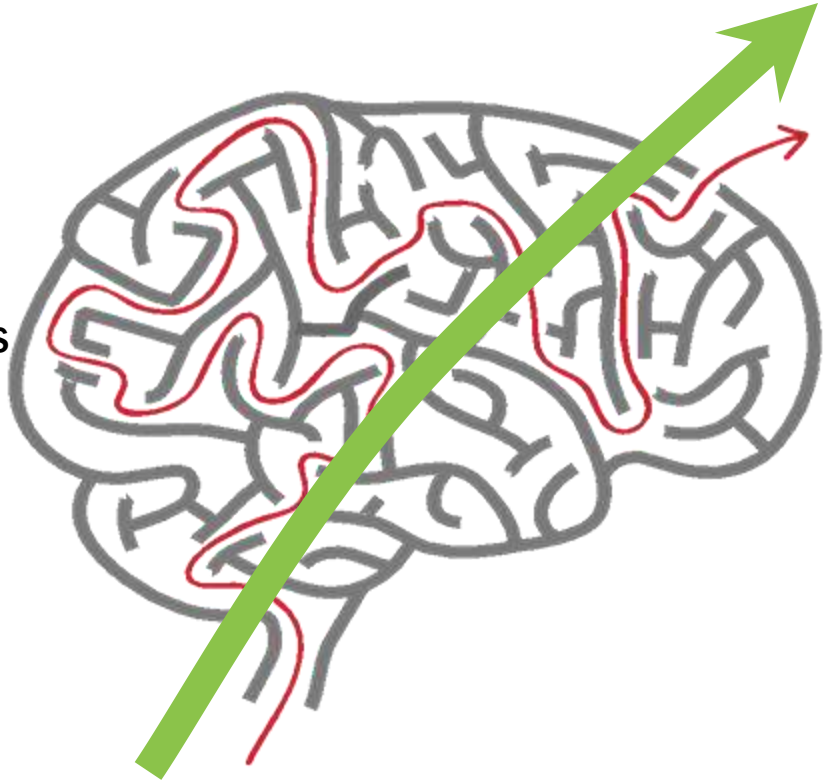
MIT Moral Machine Project



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

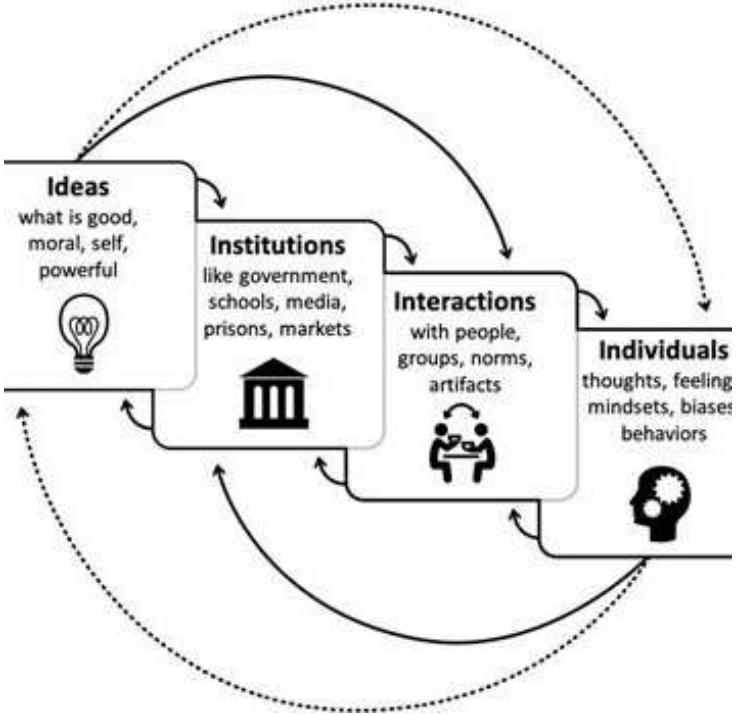
What is bias?

- Bias are shortcuts, simplify problems
- Not inherently bad, may be misapplied
- Implicit = invisible
- Not necessarily in sync with our conscious beliefs
- **Can be managed and changed**
- Talk about biases in non-threatening, productive ways



Bias due to…

- Social class
- Resource availability
- Education
- Race, gender, sexuality
- Culture, theology, tradition
- More…



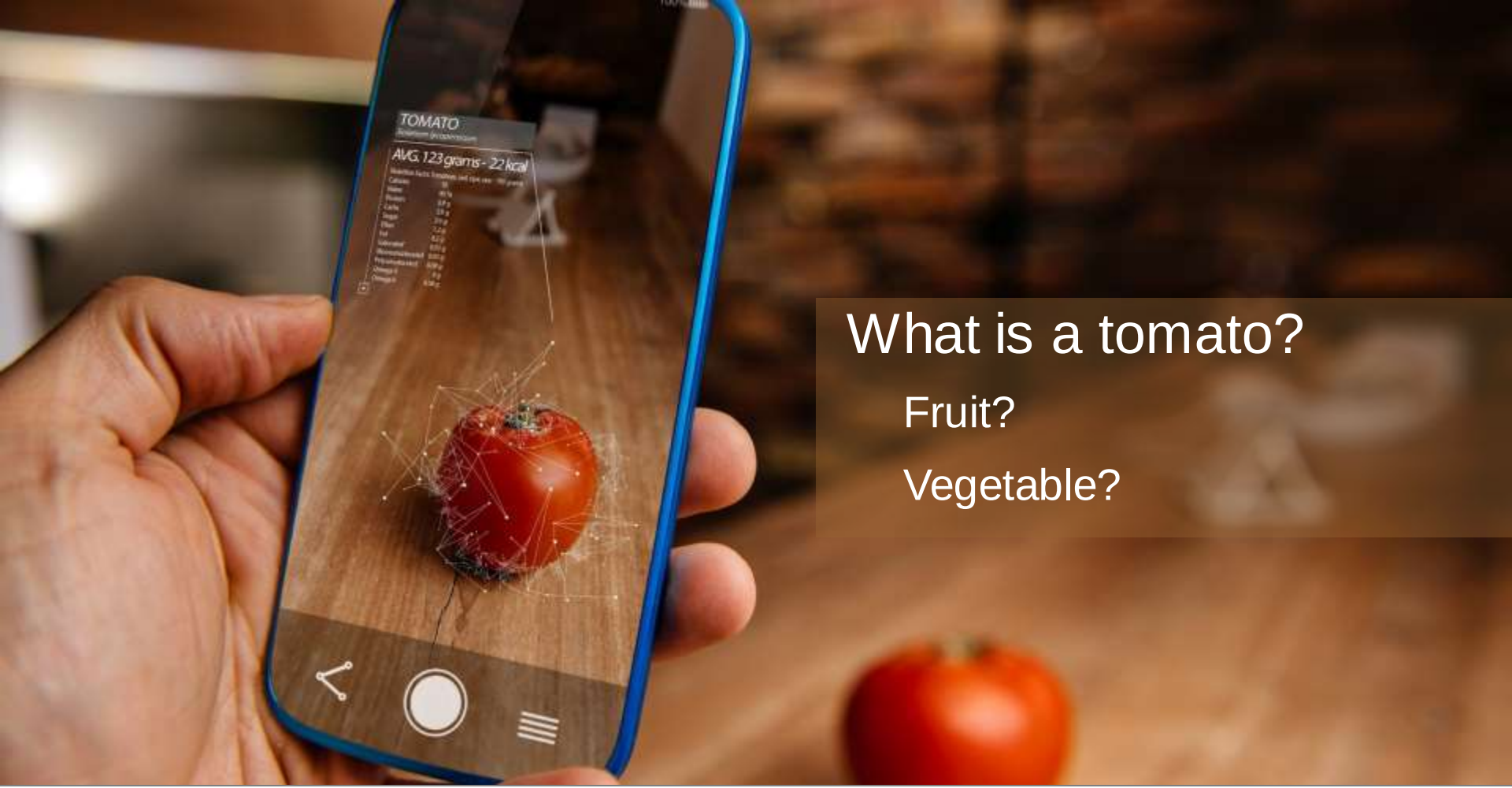
Hazel Markus, Clash!

Value diverse teams to reduce bias

- Focus more on facts and process facts more carefully
- Proactive in addressing bias
- More innovative and make more evidence-based decisions

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

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



What is a tomato?

Fruit?

Vegetable?

Bias in Image Recognition

Training data

Data encountered

Use case courtesy of Dr. Eric Heim, CMU SEI https://resources.sei.cmu.edu/library/author.cfm?authorid=542374

Only know what taught

Training data

Unrepresentative
or incomplete training data

Data encountered

Unlikely to recognize

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

Dr. Joy Buolamwini
Algorithmic Justice League
Movie: Coded Bias on Netflix

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

All systems have bias

- Complete objectivity is misleading
- Bias can have purpose and can be helpful
- The goal is to *reduce* unintended and/or harmful bias

"We often have no way of knowing when and why people are biased."
–Sandra Wachter

The Apple Card Didn't 'See' Gender—and That's the Problem
https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

Mitigating bias

Data is the primary source of bias in AI systems.

Avoid proliferation of harmful bias by understanding

- Bias in the data (purposeful and unintended, potentially harmful bias)
- Data creator's motivation and collection process
- Rationale for data inclusion, and what was excluded
- Recommended uses, etc.

Algorithm and training selection can add further bias.

Where to Start with AI

Align on technical ethics

NETWORKS / CYBER, PENTAGON

Ethical AI For War? Defense Innovation Board Says It Can Be Done

The military figured out how to use nuclear power safely, the advisory panel said, and it can do the same with artificial intelligence.

By SYDNEY J. FREEDBERG JR. on October 31, 2019 at 12:21 PM

Textron Ripsaw M5 robot in an armed configuration

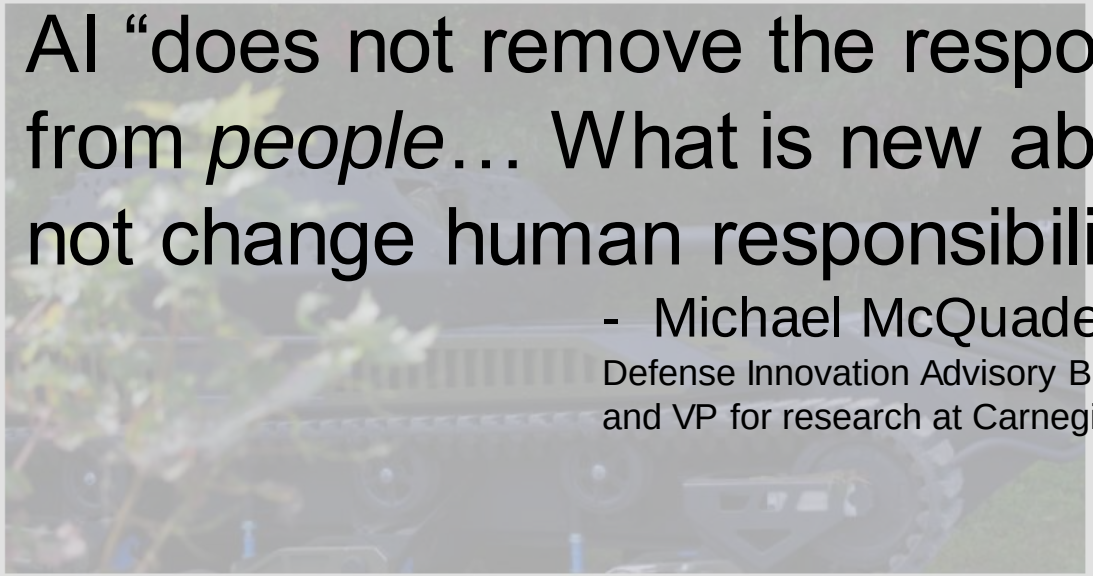
Recommended

NETWORKS / CYBER, PENTAGON

Ethical AI For War? Defense Innovation Board Says It Can Be Done

The military figured out how to use nuclear power safely, the advisory panel said, and it can do the same with artificial intelligence.

By SYDNEY J. FREEDBERG JR. on October 31, 2019 at 12:21 PM



Textron Ripsaw M5 robot in an armed configuration

AI "does not remove the responsibility from *people*… What is new about AI does not change human responsibility."

- Michael McQuade

Defense Innovation Advisory Board member
and VP for research at Carnegie Mellon University

NAVAL WARFARE, SPONSORED

When Innovation [illegible] Succeeds

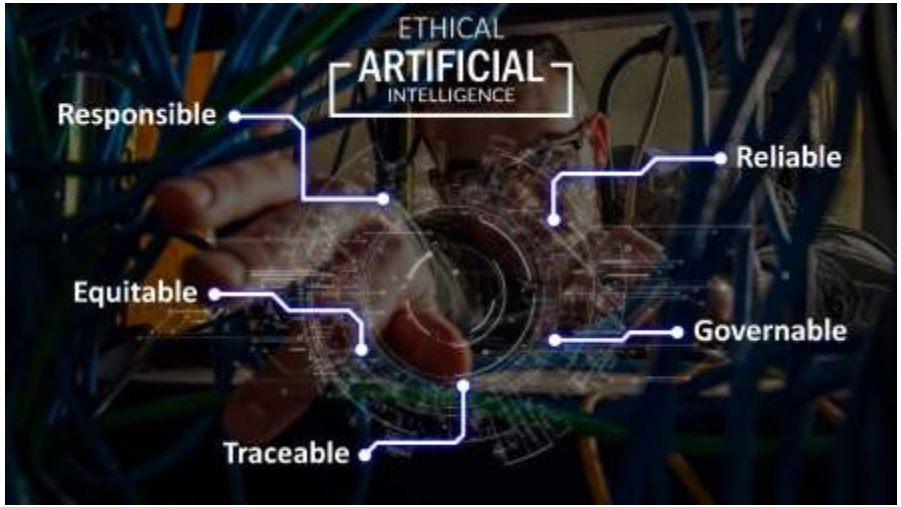
Raytheon's Naval Strike Missile will play an important role for the U.S. Navy. Learn how the USS Gabrielle Giffords became the first navy littoral combat ship to launch the NSM in an integrated setup.

From RAYTHEON

Recommended

Department of Defense Ethical Principles for Artificial Intelligence

- Responsible
- Equitable
- Traceable
- Reliable
- Governable



“Ethical considerations are an inseparable part of research, design, and deployment for DoD AI systems.” – Defense Innovation Board

Content adapted from Department of Defense Ethical Principles for Artificial Intelligence.
Department of Defense Adopts Ethical Principles for Artificial Intelligence - FEB. 24, 2020
https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/

Who are technical ethics for?

Everyone creating an AI system

- Data Scientists and data creators
- Curiosity experts
  (user experience researchers, human-computer interaction practitioners,
  sociotechnical researchers, etc.)
- Product Managers
- Machine learning experts
- Programmers, system architects

Share understanding

Converse on 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 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

Integrate technical ethics

Pair DoD Ethical Principles for AI (or another set) with responsible AI frameworks and tools

- Reduce risk and unwanted bias
- Support inspection and mitigation planning


Carnegie Mellon University
Software Engineering Institute

Designing Ethical AI Experiences: Checklist and Agreement

USE THIS DOCUMENT TO GUIDE THE DEVELOPMENT of accountable, de-risked, respectful, secure, honest, and usable artificial intelligence (AI) systems with a diverse team aligned on shared ethics. An initial version of this document was presented with the paper *Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development* by Carol Smith, available at https://arxiv.org/abs/1910.03515.

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

- ☐ Designated humans have the ultimate responsibility for all decisions and outcomes
  - Responsibilities are explicitly defined between the AI system and human(s), and how they are shared
  - Human responsibility will be preserved for final decisions that affect a person's life, quality of life, health or reputation
  - Humans are always able to monitor, control, and deactivate systems
- ☐ Significant decisions made by the AI system will be:
  - explained
  - able to be overridden
  - appealable and reversible

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

- ☐ Harmful, malicious use and consequences, as well as good, beneficial use and consequences
- ☐ We will be cognizant and exhaustively research unintended consequences

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

- ☐ communication plans to share pertinent information with all affected people
- ☐ mitigation plans for managing the identified speculative risks

We value respect and security:

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

We value transparency with the goal of engendering trust:

- ☐ The purpose, limitations, and biases of the AI system are explained in plain language.
- ☐ Data sources have unambiguous, respected sources, and biases are known and explicitly stated.
- ☐ Algorithms and models are appropriate and verifiable.
- ☐ Confidence and consent are processed for humans to base decisions on.
- ☐ Transparent justification for recommendations and outcomes is provided.
- ☐ Straightforward and interpretable monitoring systems are provided.

We value honesty and usability:

- ☐ Humans can easily discern when they are interacting with the AI system vs. a human.
- ☐ Humans can easily discern when and why the AI system is taking action and/or making decisions.
- ☐ Improvements will be made regularly to meet human needs and technical standards.

Team Signatures and Date

About the SEI

Contact Us

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

Activate curiosity

Incorporate user experience (UX) research and human-computer interaction (HCI) methods to activate curiosity.

- Speculate about misuse and abuse.
- Identify potential unintended/unwanted consequences.
- Potential severe abuse and consequences.
- Perspectives of people in frequently marginalized groups.

Where to Start with AI

Managing AI

AI systems have great potential, and require forethought

- Determine ownership of data, models, etc. in contracts
- Plan for long term implementation and oversight
- Prevent harm, and plan for mitigation
- AI systems
  - Require significant AI-appropriate data
  - Need to be verified and validated frequently
  - Are not “stable” like typical software
  - Are not an inexpensive solution

Responsible AI systems should be…

- Accountable to humans
- Built speculatively
- Respectful and secure
- Honest and usable

Check in with
"What are we doing?
Why are we doing it,
and for whom?"

Nacho Kamenov & Humans in the Loop / Better Images of AI /
A trainer instructing a data annotator on how to label images / CC-BY 4.0

Design human-centered AI systems

AI systems must be designed so that humans

- Can monitor and control risk.
- Are always ultimately in control, and ultimately responsible for all decisions and outcomes.
- **Are responsible for final decisions** that affect a person's life, quality of life, health, or reputation (not an AI).

Challenges

- State of the art in AI technology is continuously evolving
- Intended level of interdependence between humans and machines leads to trust and transparency challenges
- AI systems will be deployed alongside humans in unfamiliar and unpredictable operational contexts
- Guidelines and heuristics are scarce for understanding and implementing the level of oversight needed to create and maintain ethical AI systems

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

Support teams in making effective and responsible AI systems.

“AI will ensure appropriate human judgement and not replace it” - DIB

AI in the World – Bad Examples

Dangerous behavior

Cruise driverless vehicles block San Francisco traffic, face regulatory scrutiny

Print

By Brad Matthews - The Washington Times - Wednesday, September 28, 2022

California regulators said Tuesday they are examining multiple incidents of stalled Cruise driverless vehicles blocking traffic in San Francisco last week.

The Washington Times

Dehumanizing thru Math

Predicting future criminal activity

Embedding Existing Bad Behaviors

Determining interest rates for mortgage lenders

AI-generated images

An A.I.-Generated Picture Won an Art Prize

Jason Allen's A.I.-generated work, "Théâtre D'opéra Spatial," took first place in the digital category at the Colorado State Fair. via Jason Allen

AI in the World – Good Examples

Quality Control

A vision system is trained to recognize a ‘bad’ image versus a ‘good’ image for bottle labels. Courtesy of Artemis Vision.

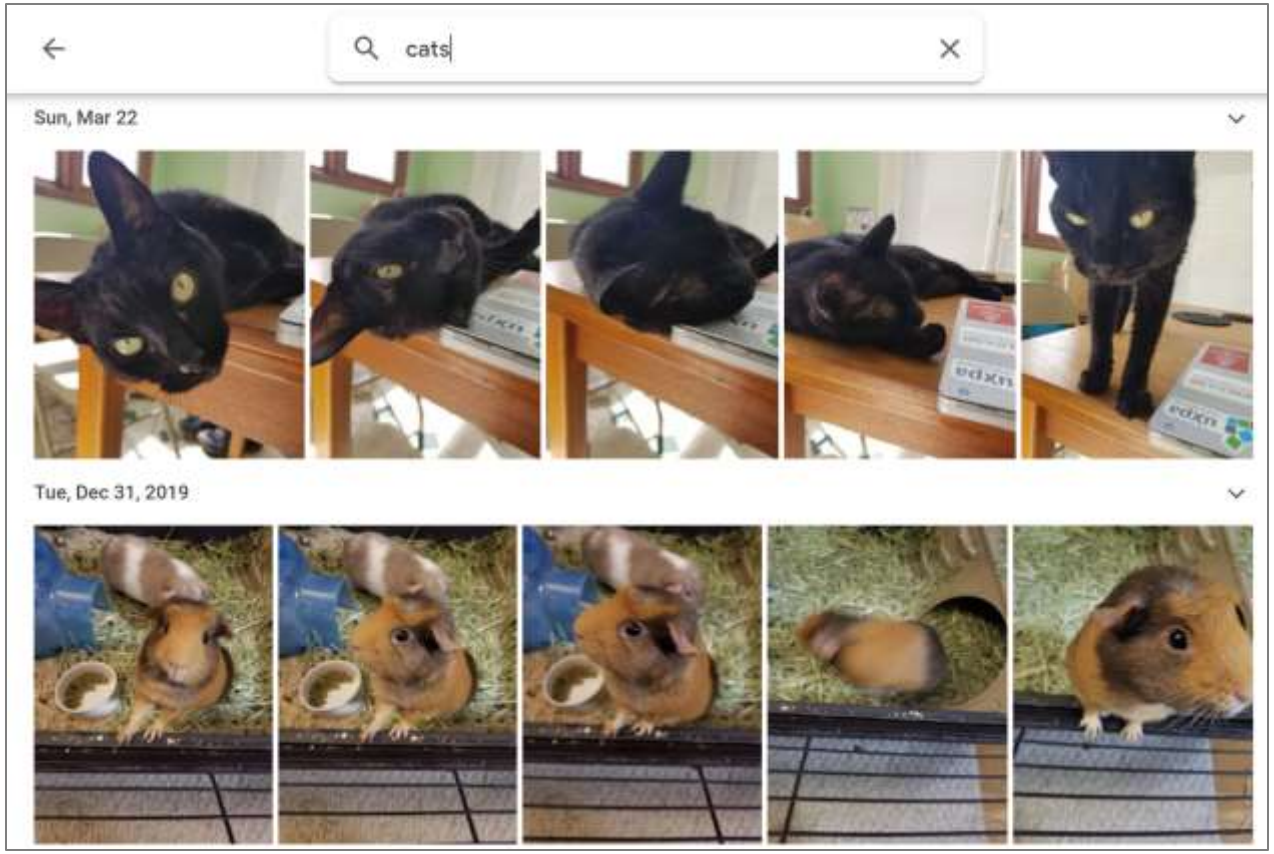
Strategic Games

1997 Chess, IBM

2016 Go, Google

Floor goban, 2007, By Goban1
https://commons.wikimedia.org/wiki/File:FloorGoban.JPG

Image Recognition – Google Photos



Carol's search for "cats" on her Google Photos account.

Sound recognition: Labeling of birdsongs

"Comparison of machine learning methods applied to birdsong element classification"
by David Nicholson. Proceedings of the 15th Python in Science Conference (SCIPY 2016). http://conference.scipy.org/proceedings/scipy2016/pdfs/david_nicholson.pdf
Photo by Gallo71 (Own work) [Public domain], via Wikimedia Commons https://commons.wikimedia.org/wiki/File%3ABrun.JPG

Listening and understanding human speech

Mapping Q & A + AI

- Expected language
- Appropriate automated responses
- When to escalate?
  - Searches on self harm?
  - What else?

Hi, I'm Woebot|

Images: https://www.pexels.com/photo/close-up-of-mobile-phone-248512/
https://www.amazon.com/Amazon-Echo-Bluetooth-Speaker-with-WiFi-Alexa/dp/B00X4WHP5E
https://www.ibm.com/watson/developercloud/doc/conversation/index.html

Thank You

Alex Van Deusen
Design Researcher, CMU SEI
arvandeusen@sei.cmu.edu

Carol J. Smith
Sr. Research Scientist, CMU SEI
cjsmith@sei.cmu.edu

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213