Intro to Ethical & Trustworthy AI

Destiny Pounds

M.S. Student, Biomedical Data Science

Advised by Vibhuti Gupta, Ph.D.

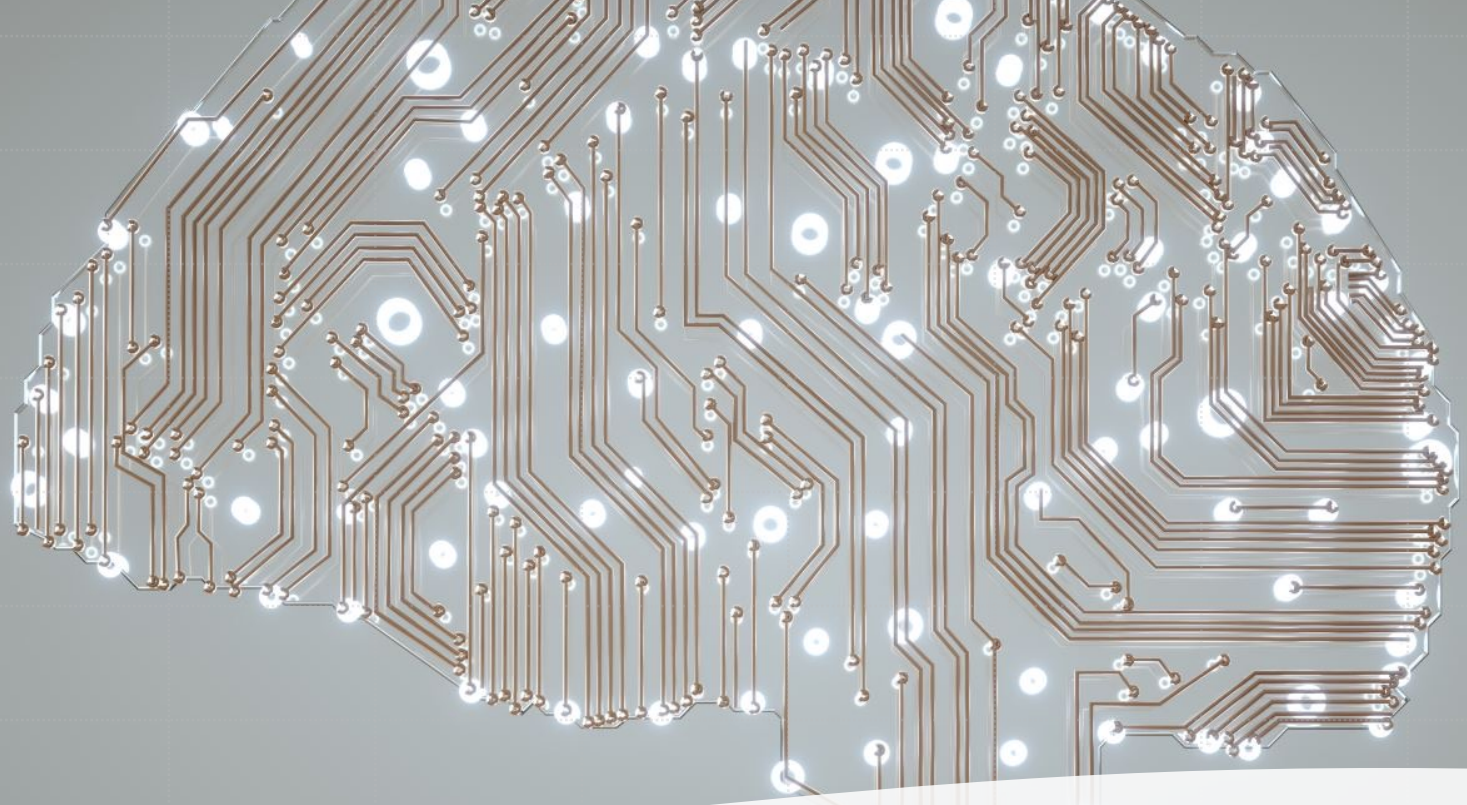
Assistant Professor, Computer Science and Data Science

School of Applied Computational Sciences

Meharry Medical College

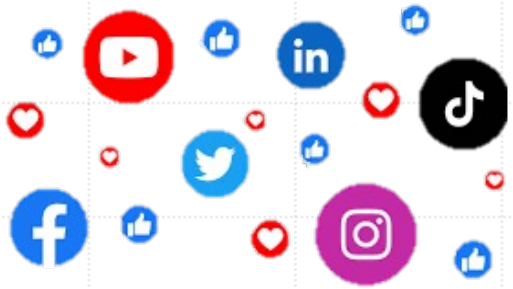
Overview

- What is Artificial Intelligence?
- Contributions & Concerns
- Machine Learning Pipeline
- Characteristics of Trustworthy AI
  - Ethics
  - Fairness
  - Privacy
  - Security
  - Robustness
  - Safety
  - Explainability
  - Accountability



Artificial Intelligence (AI) describes a program or system that can effectively address real-world problems in a human-like way.

Everyday Examples of AI Use



Personalized social media feeds

Google Maps

Traffic monitoring and route suggestion

Email filters

amazon

Shopping recommendations

The goal is to use AI systems to sustainably benefit society across different industries.

economics

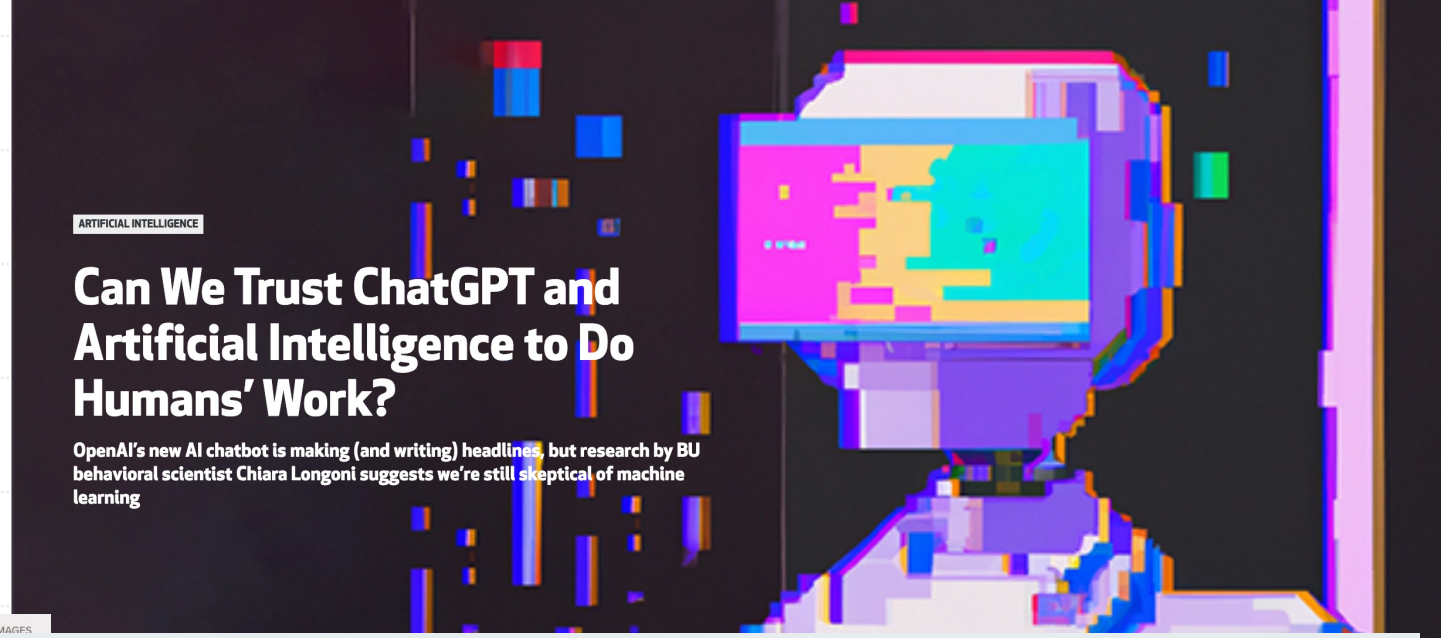
education

finance

healthcare

transportation

AI must be trustworthy to be beneficial.

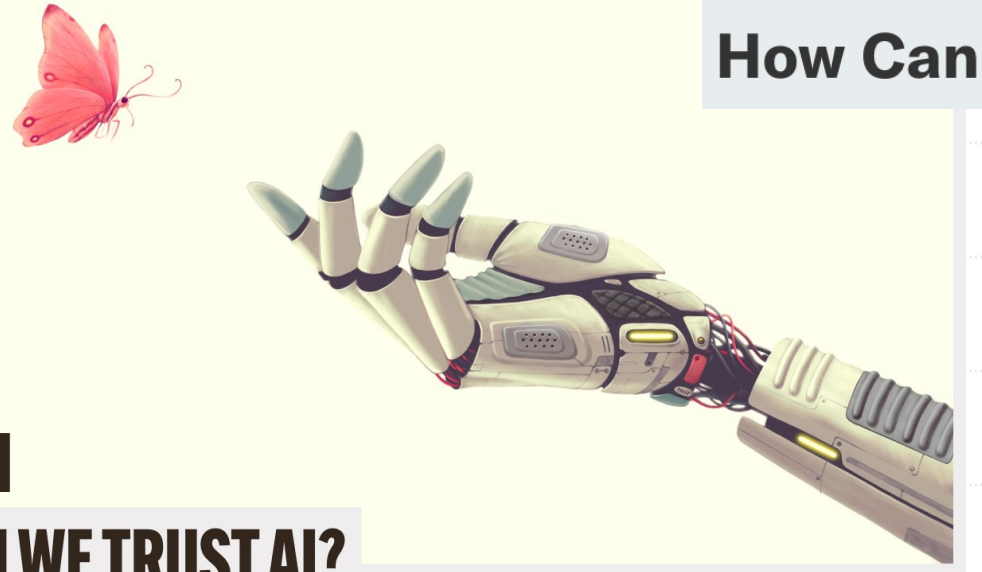


ARTIFICIAL INTELLIGENCE

Can We Trust ChatGPT and Artificial Intelligence to Do Humans' Work?

OpenAI's new AI chatbot is making (and writing) headlines, but research by BU behavioral scientist Chiara Longoni suggests we're still skeptical of machine learning

GETTY IMAGES



Q+A

CAN WE TRUST AI?

From Alexa to a robot running amok in the movie 'M3GAN', artificial intelligence is part of everyday life and is capturing our imagination. Johns Hopkins AI expert Rama Chellappa helps us sort out fact from fiction, and whether we should embrace the 'AI spring'.

OCTOBER 3, 2023 | 4 MIN READ

How Can We Trust AI If We Don't Know How It Works

Is Artificial Intelligence Good For Society?

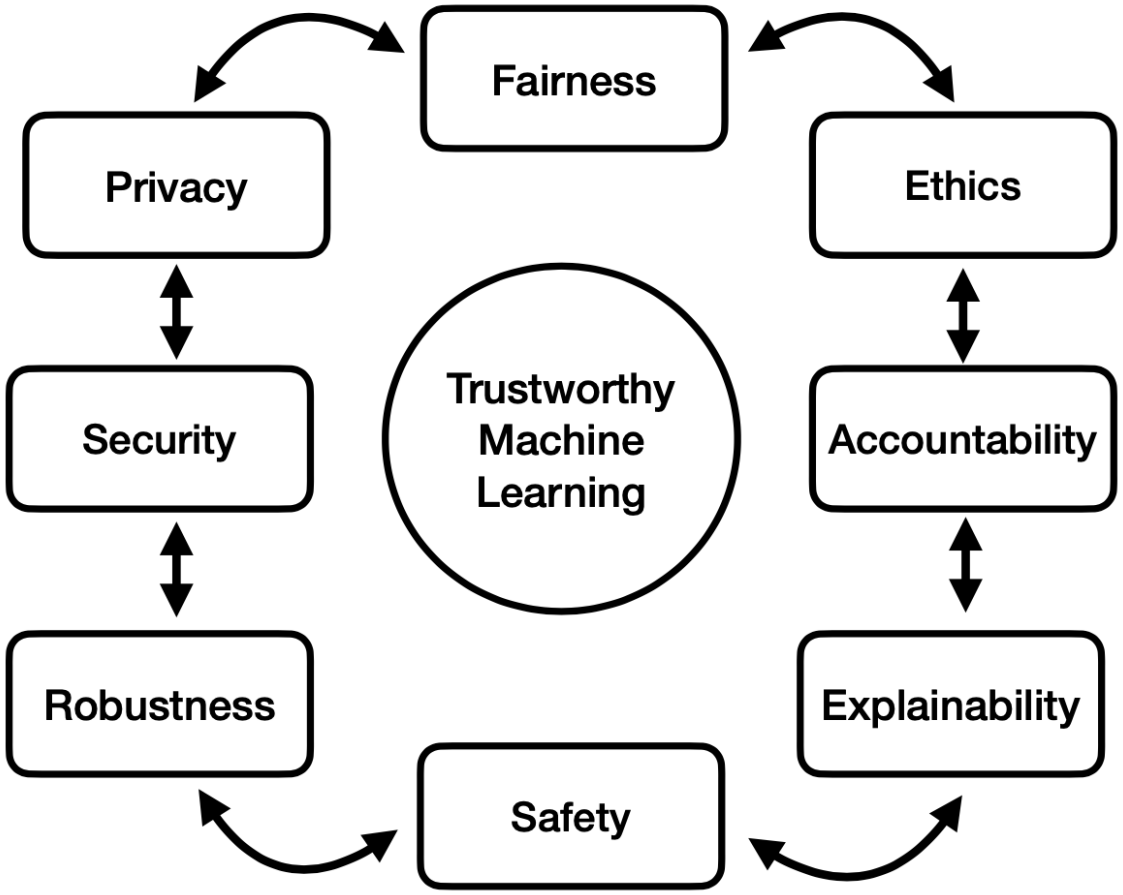
Q.ai - Powering a Personal Wealth Movement Former Contributor ⓘ

Making wealth creation easy, accessible and transparent.

0

Feb 16, 2023, 04:24pm EST

What is Trustworthy AI?



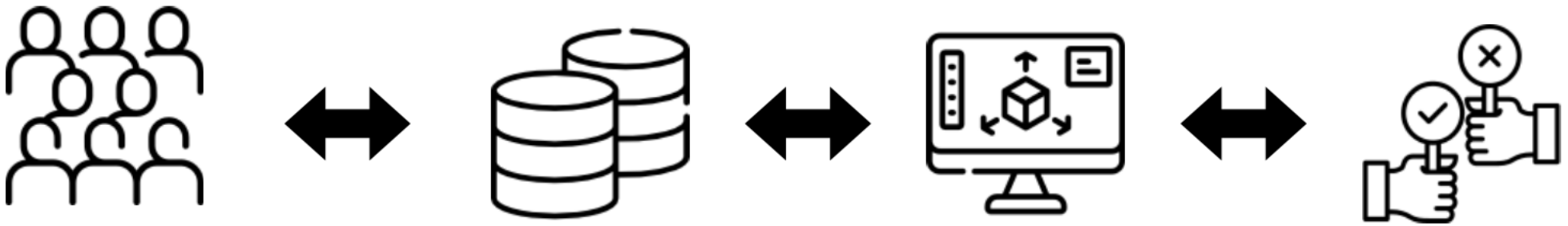
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Machine Learning Pipeline

Data is collected from individuals.

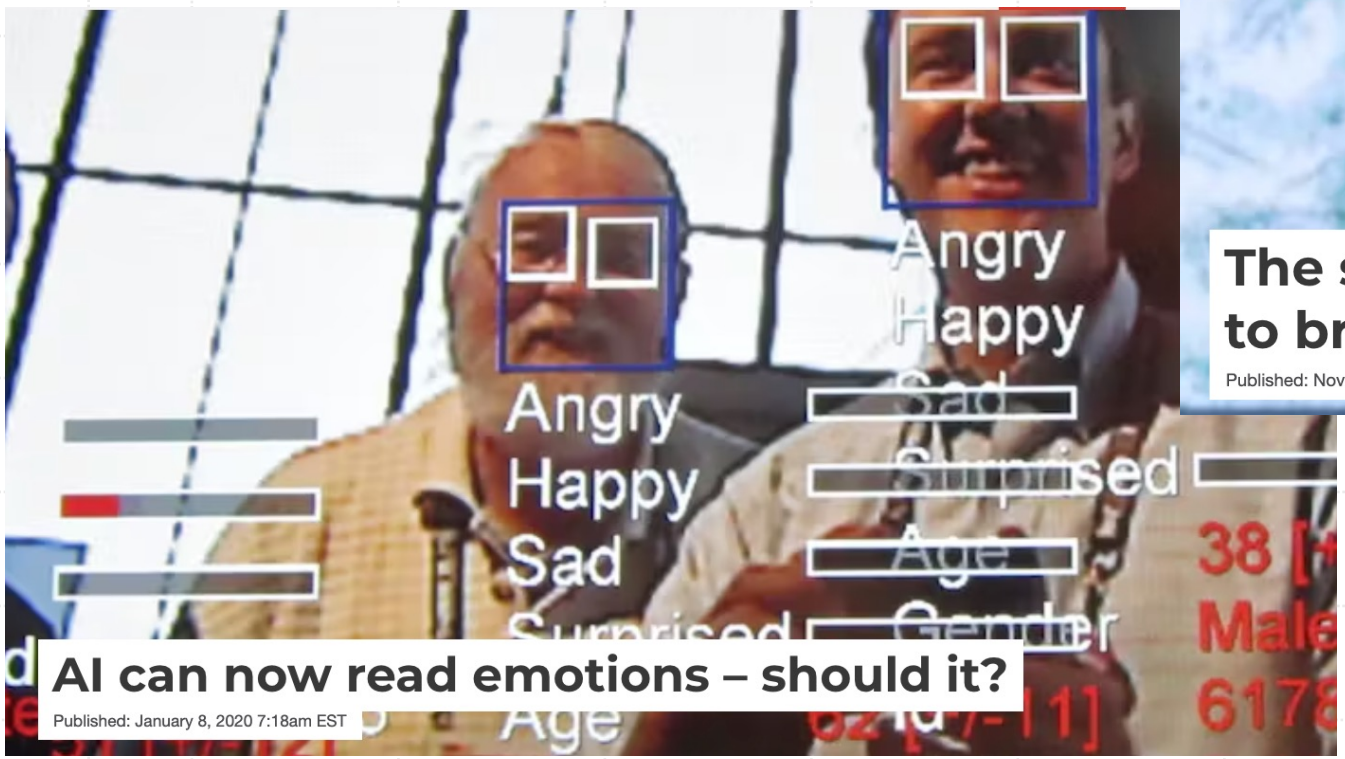
Data is used to educate AI system.

AI system makes informed decisions.



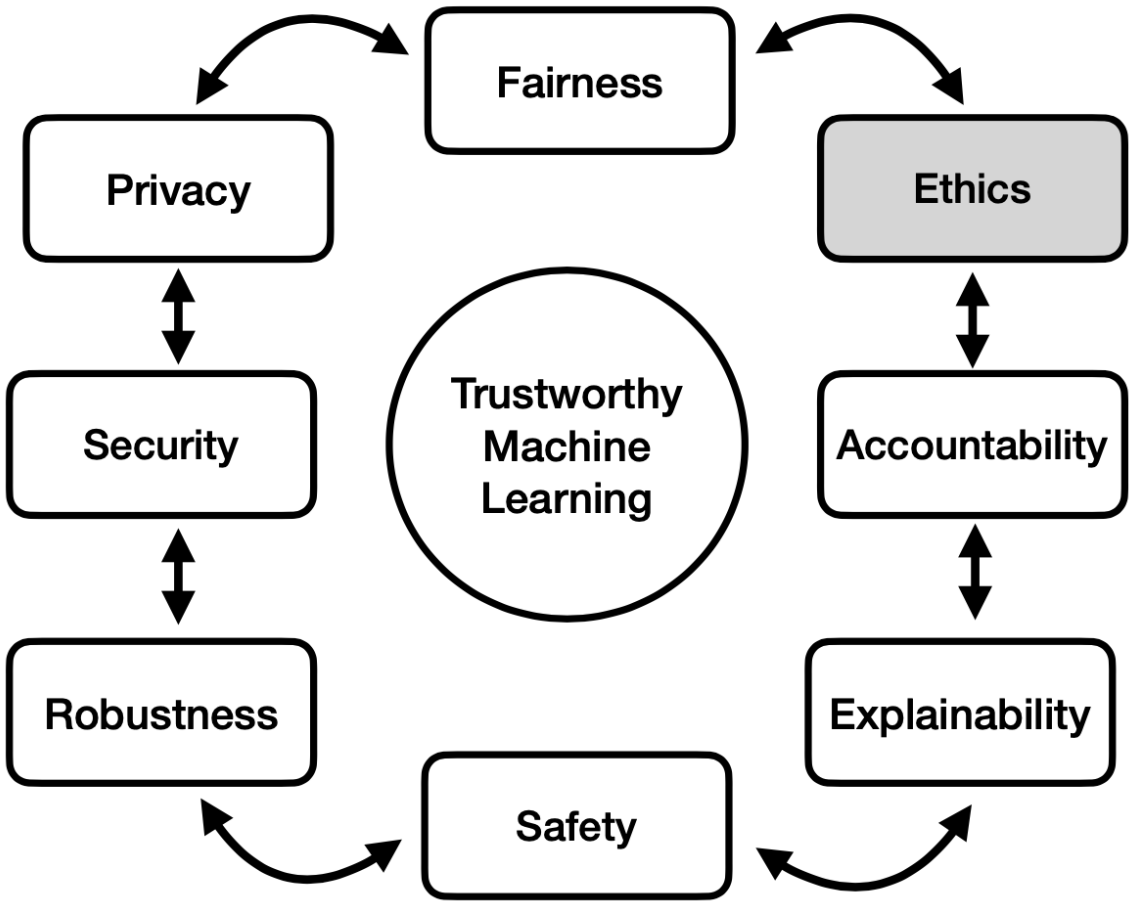
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

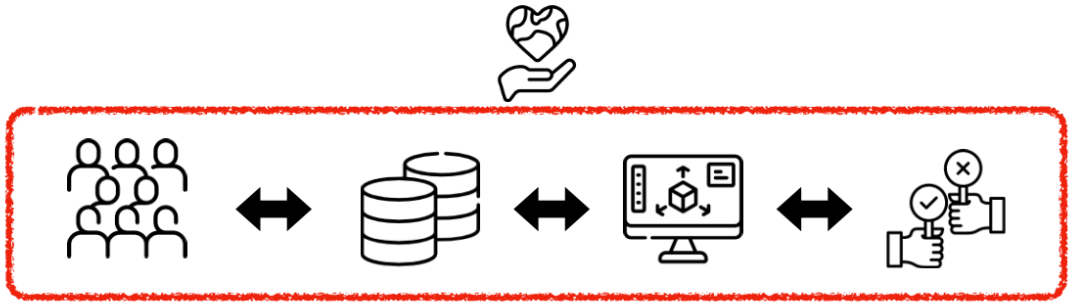
Ethics


Angry
Happy
Sad
Surprised
Age
Gender
Angry
Happy
Sad
Surprised
Age
38
Male
AI can now read emotions – should it?
Published: January 8, 2020 7:18am EST


The slippery slope of using AI and deepfakes to bring history to life
Published: November 2, 2021 8:26am EDT

Ethics





- AI ethics is a set of guidelines that advise on the design and outcomes of artificial intelligence

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

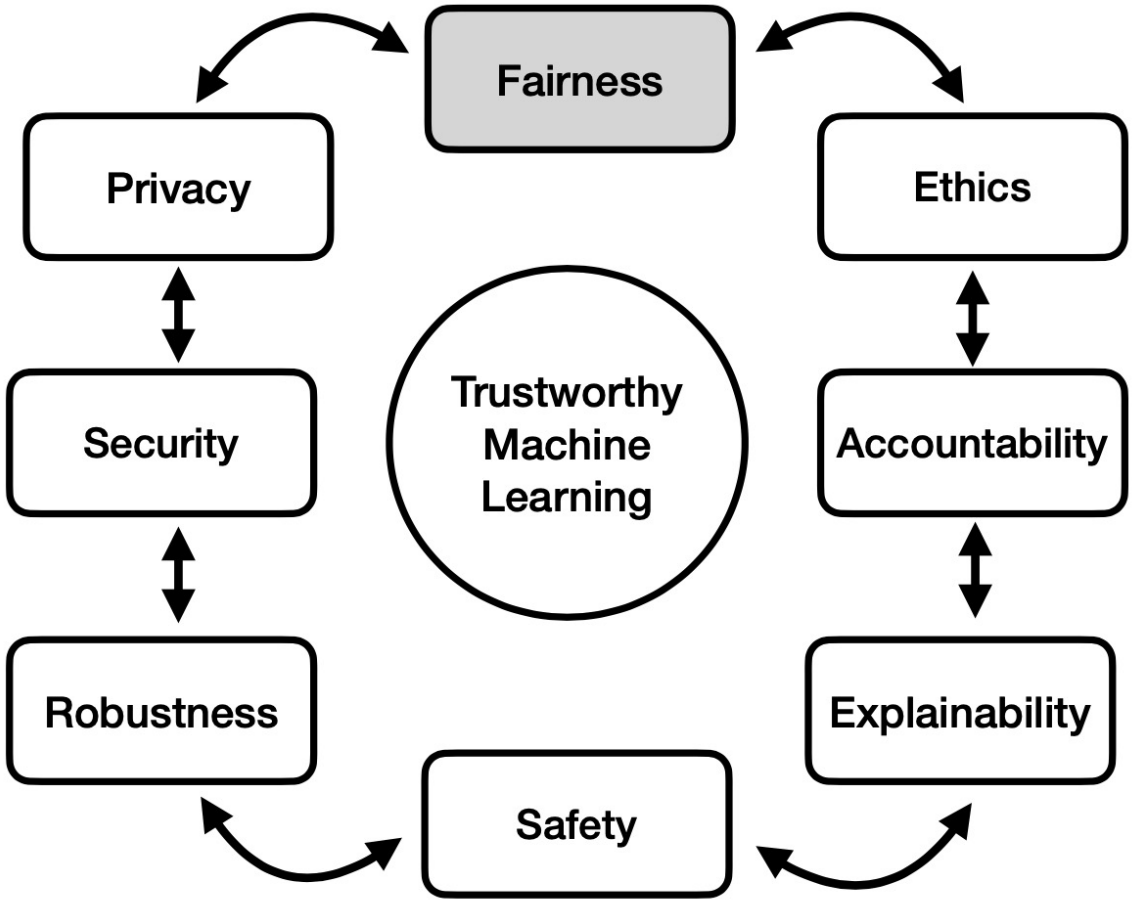
Fairness

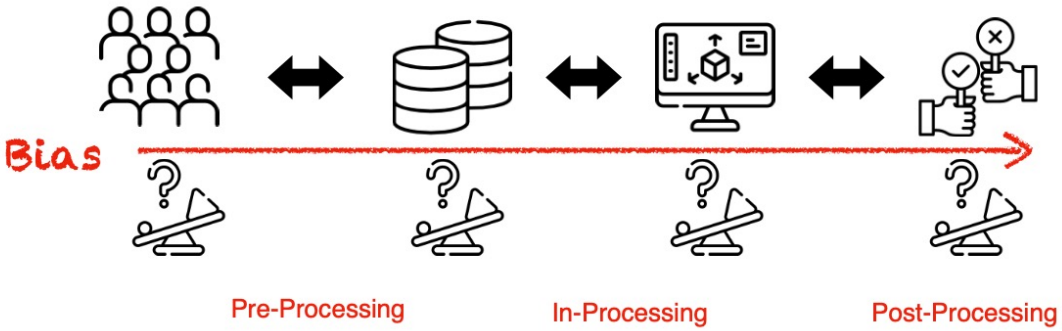
ACLU finds Amazon's facial recognition AI is racially biased

27 July 2018 | Amazon

A test of Amazon's facial recognition technology by the ACLU has found it erroneously labelled those with darker skin colours as criminals more often. Bias in AI technology, when used by law enforcement, has raised concerns of infringing on civil rights by automated racial profiling. A 2010 study by researchers at NIST and the University of Texas in Dallas found that algorithms designed and tested in East Asia are better at recognising East Asians, while those designed in Western...

Fairness





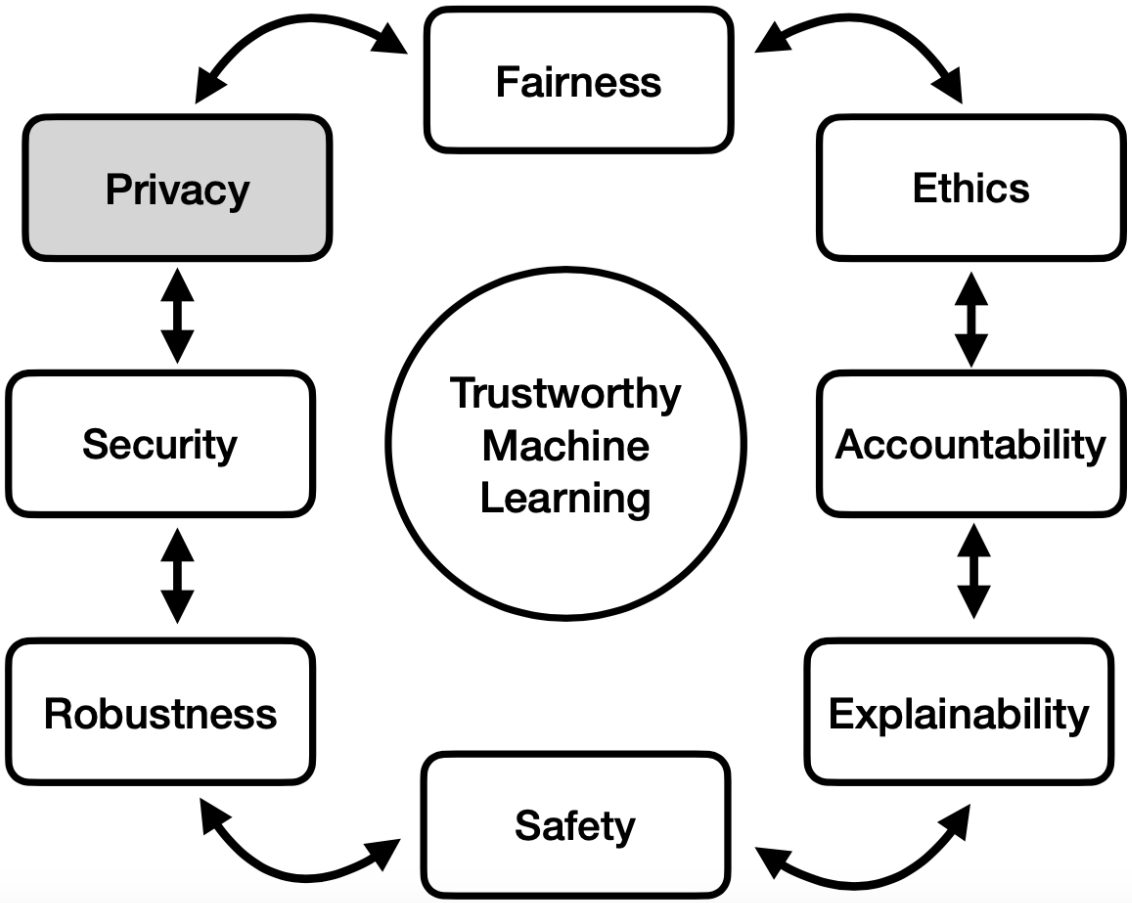
- Fairness in machine learning refers to the various attempts at correcting algorithmic bias in automated decision processes.

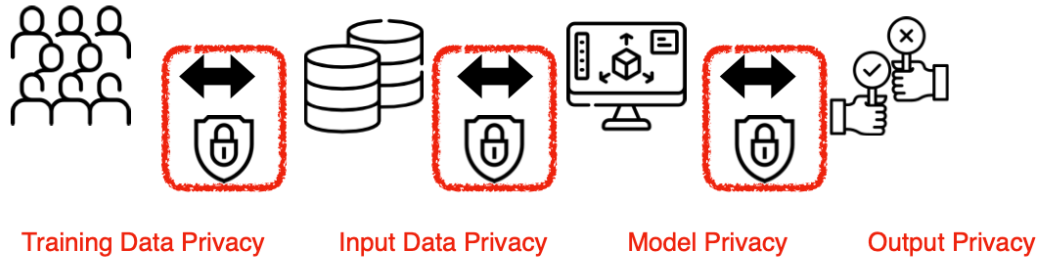
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Privacy


US agencies buy vast quantities of personal information on the open market – a legal scholar explains why and what it means for privacy in the age of AI
Published: June 29, 2023 8:16am EDT

Privacy





- Data privacy is a central issue to training and testing AI models, especially ones that train and infer on sensitive data.

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Security

HEALTH IT, MEDCITY INFLUENCERS

Hacking healthcare: Protecting patient data while maintaining access

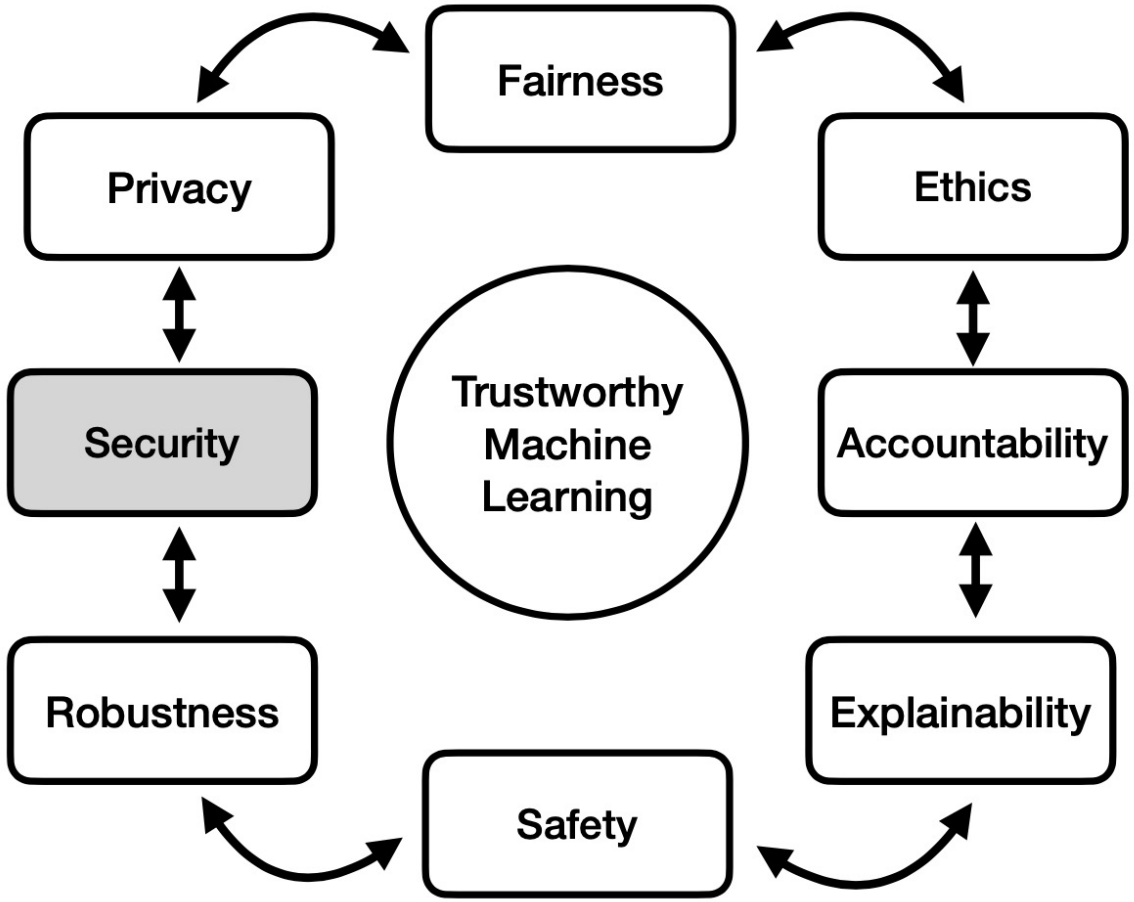
With cyber attacks on a steady incline, something needs to be done so that similar fates can be avoided and patient trust can be prioritized. There are four key best practices that, with the right data access technology in place, can protect healthcare companies from hacks and attacks

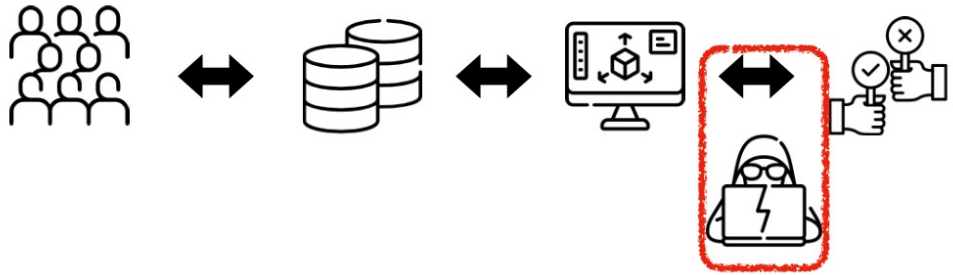
By ELDAD CHAI

Post a comment / Aug 9, 2022 at 9:00 AM

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Security





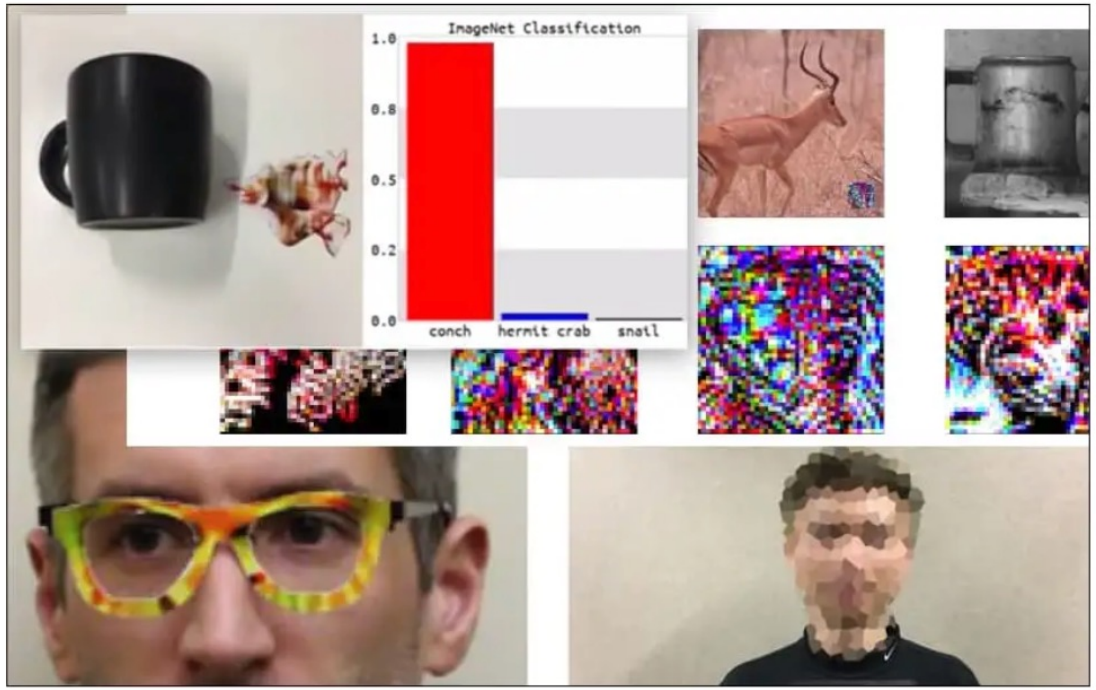
- Cybersecurity is the practice of protecting systems, networks, and programs from digital attacks.

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

CYBERSECURITY

Why Adversarial Image Attacks Are No Joke

Updated on December 1, 2021
By **Martin Anderson**



Attacking image recognition systems with carefully-crafted adversarial images has been considered an amusing but trivial proof-of-concept over the last five years. However, new research from Australia suggests that the casual use of highly popular image datasets for commercial AI projects could create an enduring new security problem.

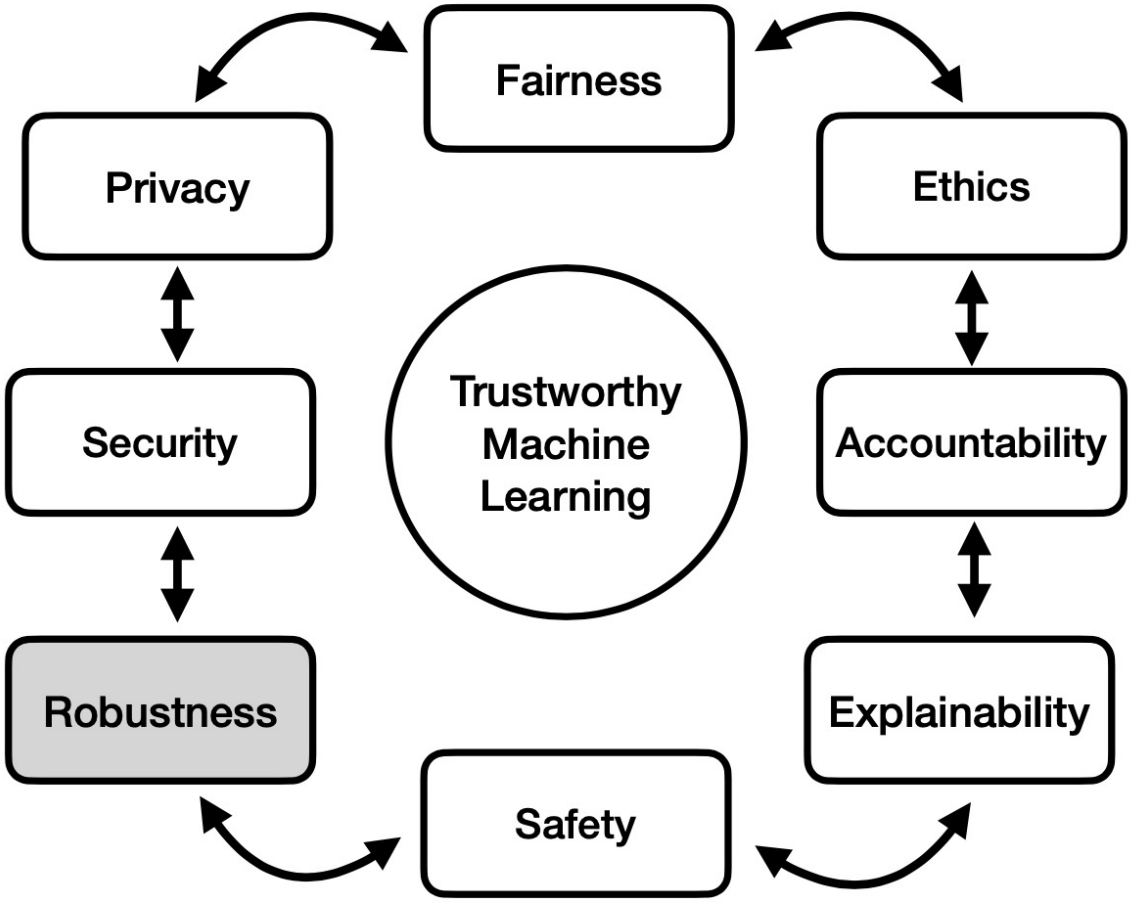
Physical adversarial example from CVPR 2018 paper

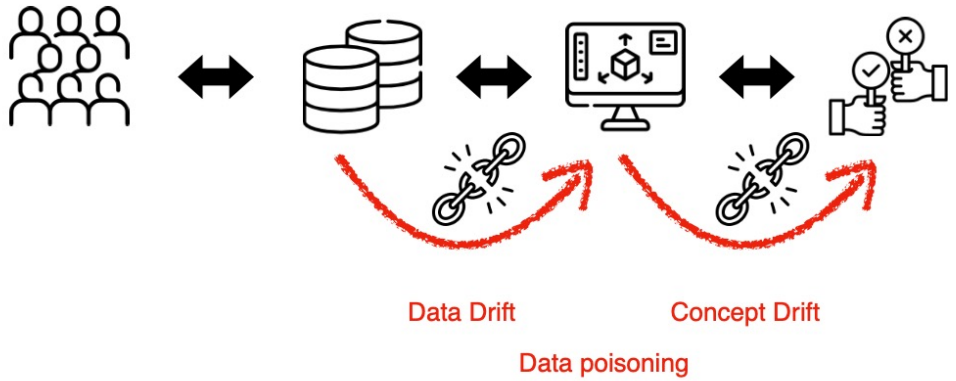
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Can we fool AI?

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Robustness





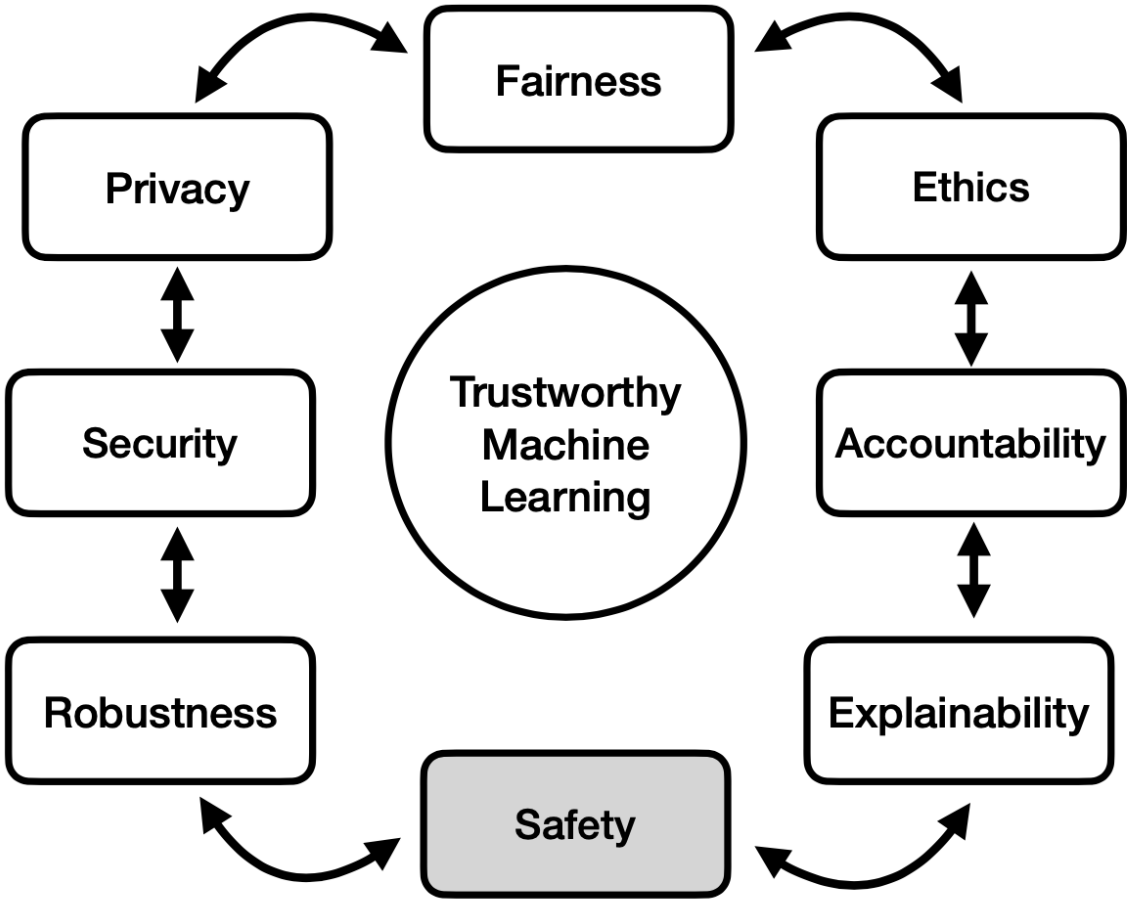
- The robustness is the property that characterizes how effective your algorithm is while being tested on the new independent (but similar) dataset.

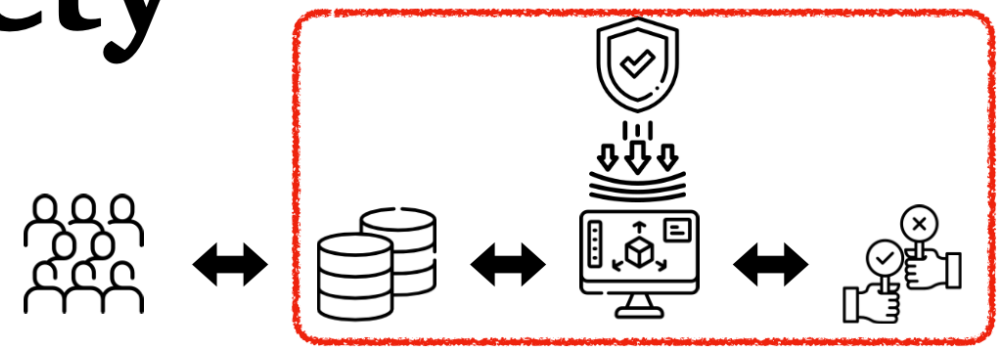
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Safety


Are autonomous cars really safer than human drivers?
Published: February 2, 2018 6:29am EST

Safety





- AI Safety can be broadly defined as the endeavour to ensure that AI is deployed in ways that do not harm humanity.
- AI Safety identifies causes of unintended behavior in machine learning systems and develop tools to ensure these systems work safely and reliably.

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Explainability

Explainable AI for Fraud Prevention

As the use of AI- and ML-driven decision-making draws transparency concerns, the need increases for explainability, especially when machine learning models appear in high-risk environments.

David Utassy
Data Scientist, SEON

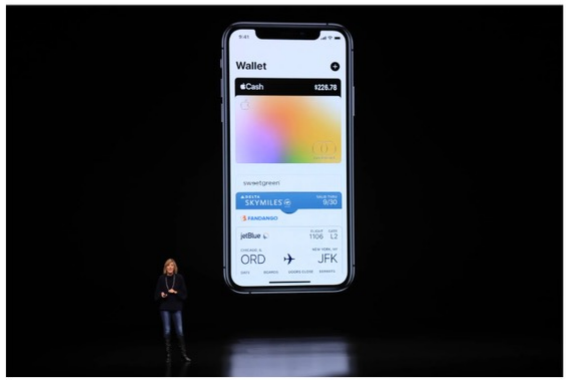
April 28, 2022

Source: Wavebreakmedia Ltd UC6 via Alamy Stock Photo

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was "sexist" against women applying for credit.

Give this article



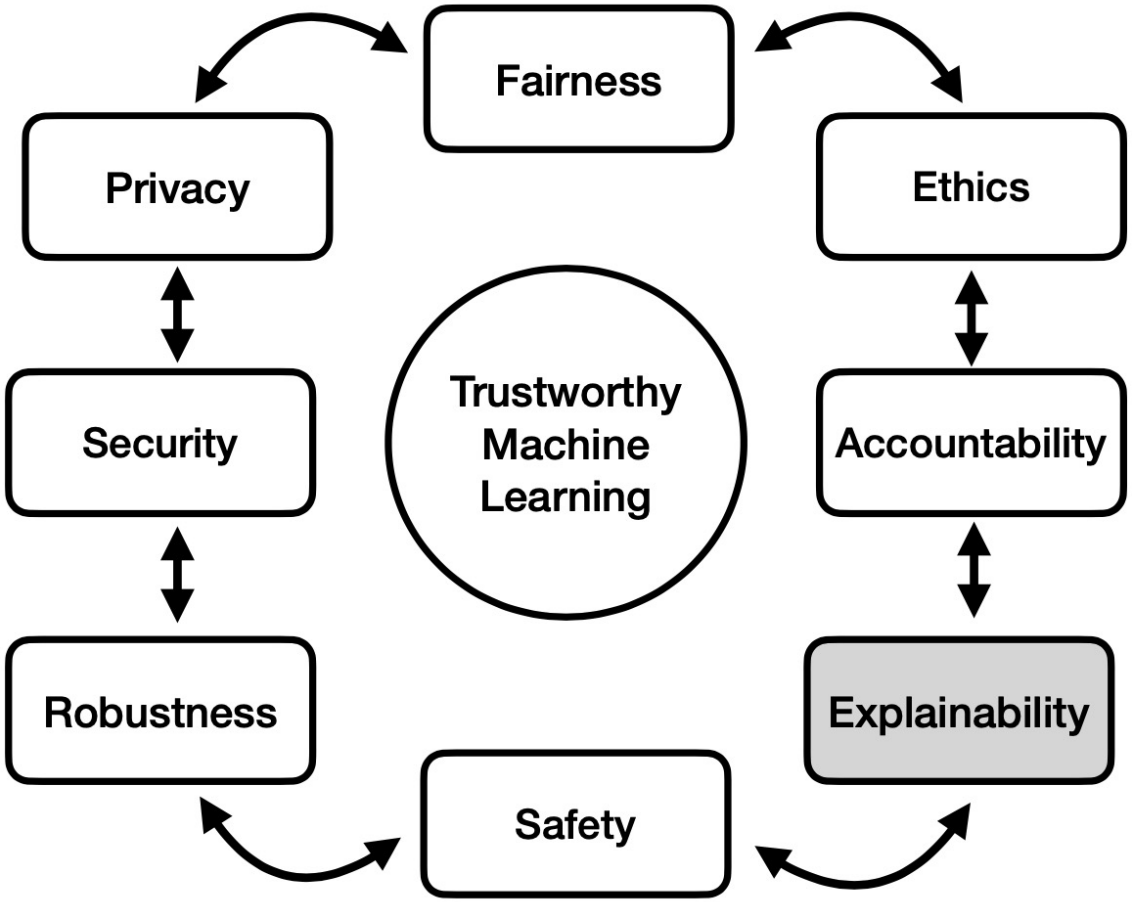
Jennifer Bailey, vice president of Apple Pay. Regulators are investigating Apple Card's algorithm, which is used to determine applicants' creditworthiness. Jim Wilson/The New York Times

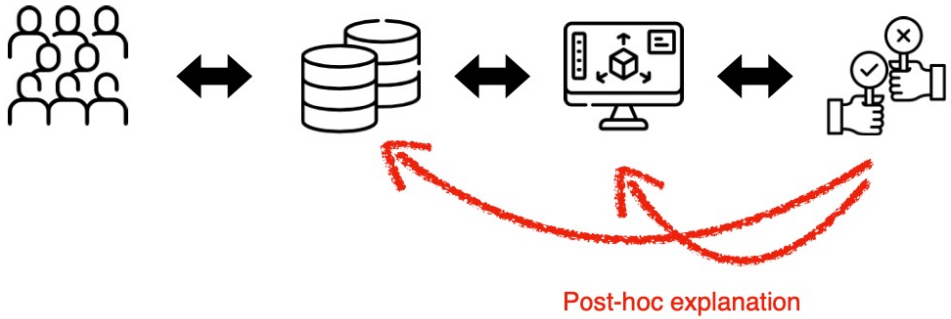
By Neil Vigdor

Nov. 10, 2019

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Explainability



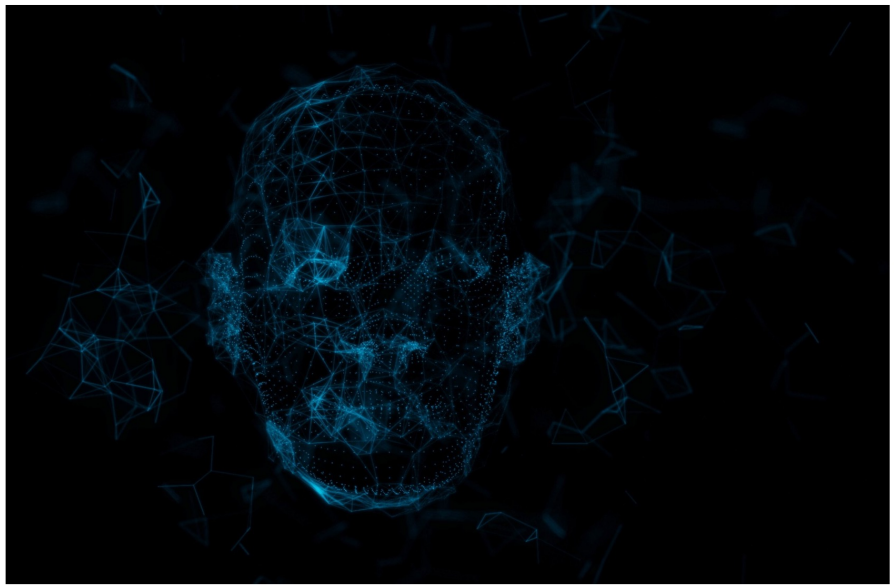


- Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms.

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Accountability

Error-prone facial recognition leads to another wrongful arrest



About the Author

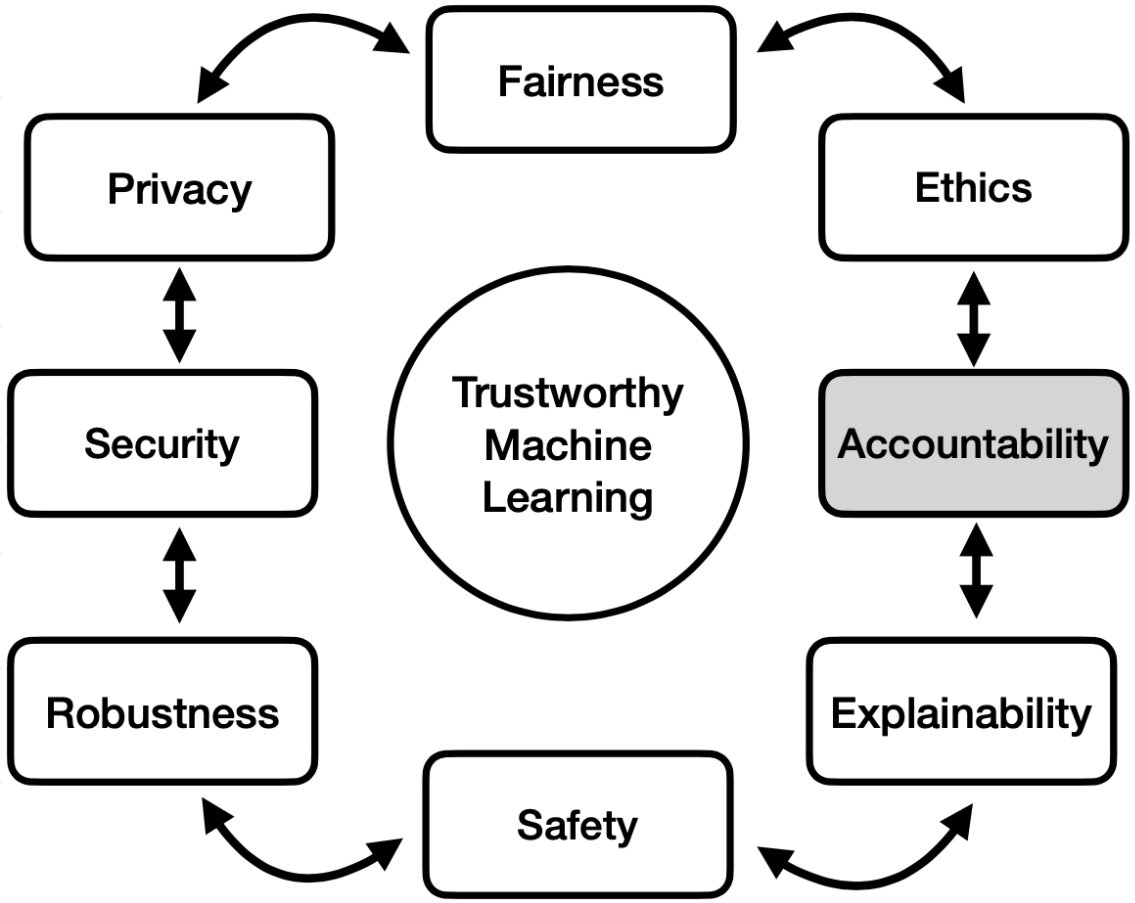
By Ryan Daws | August 7, 2023
Categories: Applications, Artificial Intelligence, Ethics & Society, Face Recognition, Privacy, Surveillance,

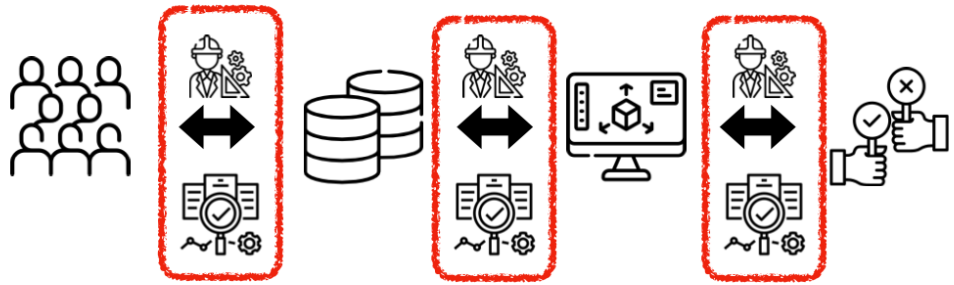
Social media algorithms are still failing to counter misleading content

About the Author

By Ryan Daws | August 17, 2021
Categories: Ethics & Society, Machine Learning, Meta (Facebook),

Accountability





- Accountability is defined as being able to ascertain whether an AI system is behaving as promised, which is necessary for determining blame-worthiness.

Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Thank You

Please send us your questions at:

vgupta@mmc.edu and

dpounds24@email.mmc.edu