Intro to Ethical & Trustworthy AI

Destiny Pounds

M.S. Student, Biomedical Data Science

Advised by Vibhuti Gupta, Ph.D.

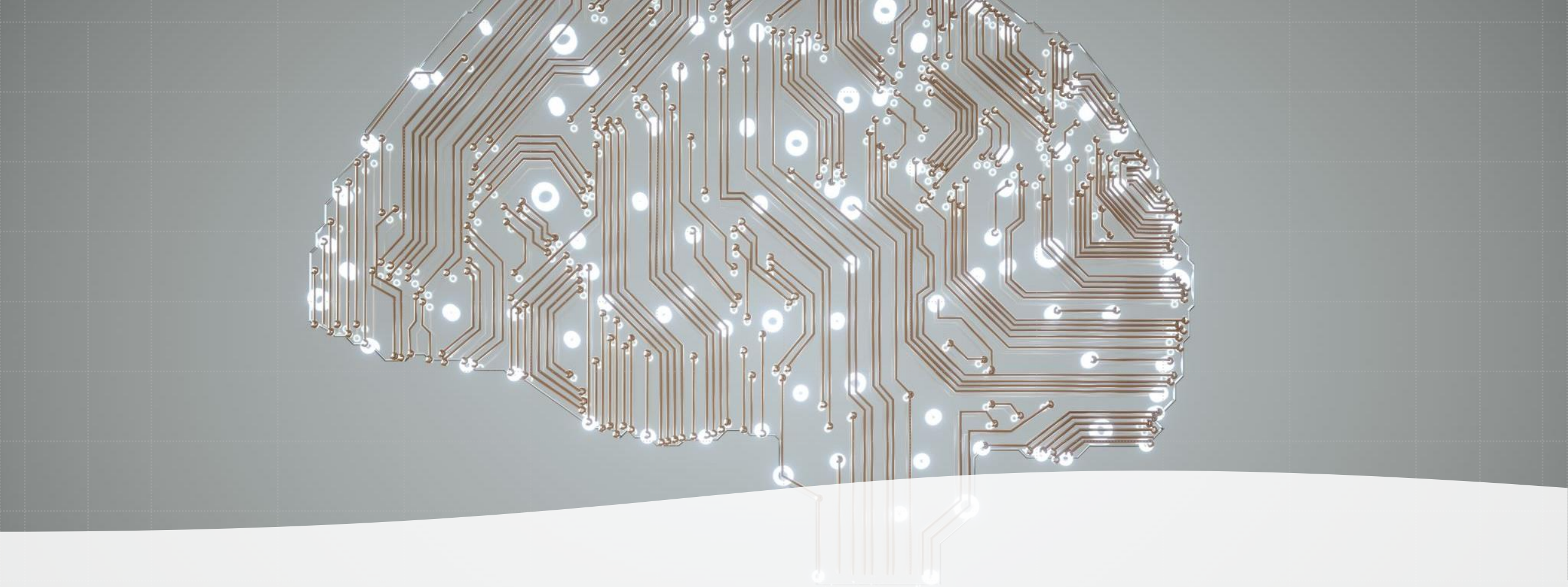
Assistant Professor, Computer Science and Data Science

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

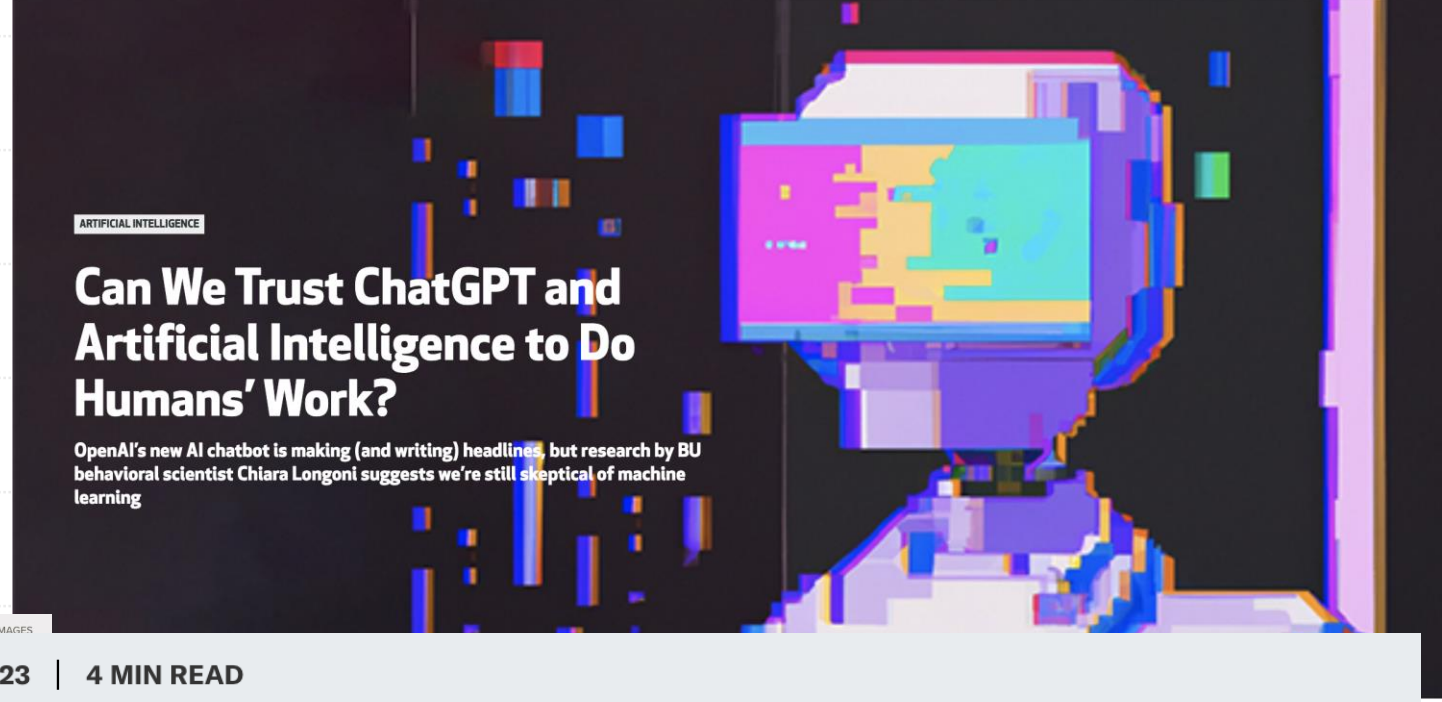
education

finance

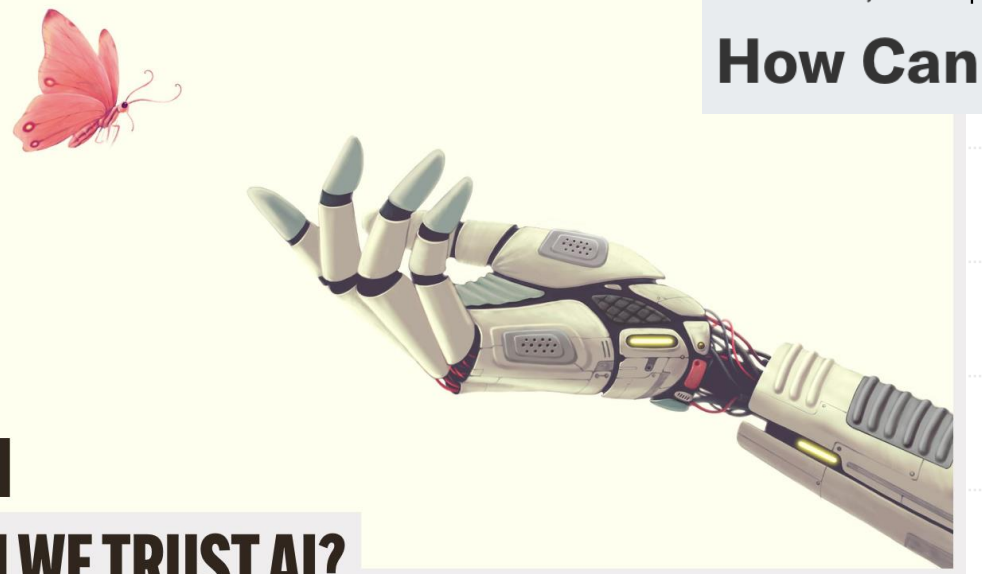
healthcare

transportation

AI must be trustworthy to be beneficial.



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

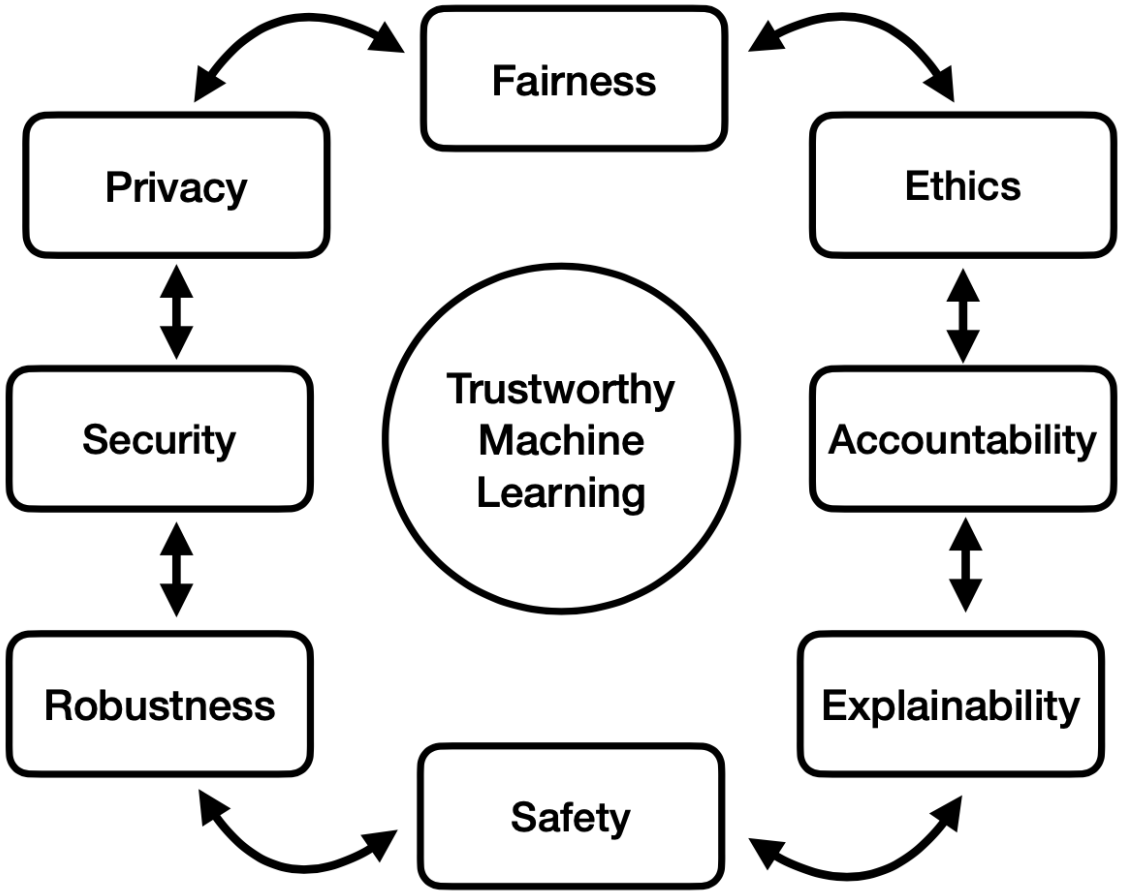
Is Artificial Intelligence Good For Society?

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?

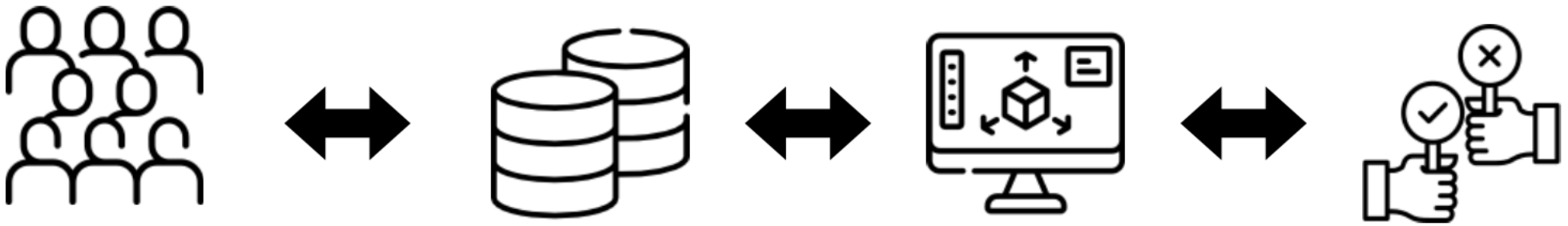
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Machine Learning Pipeline

Data is collected from individuals.

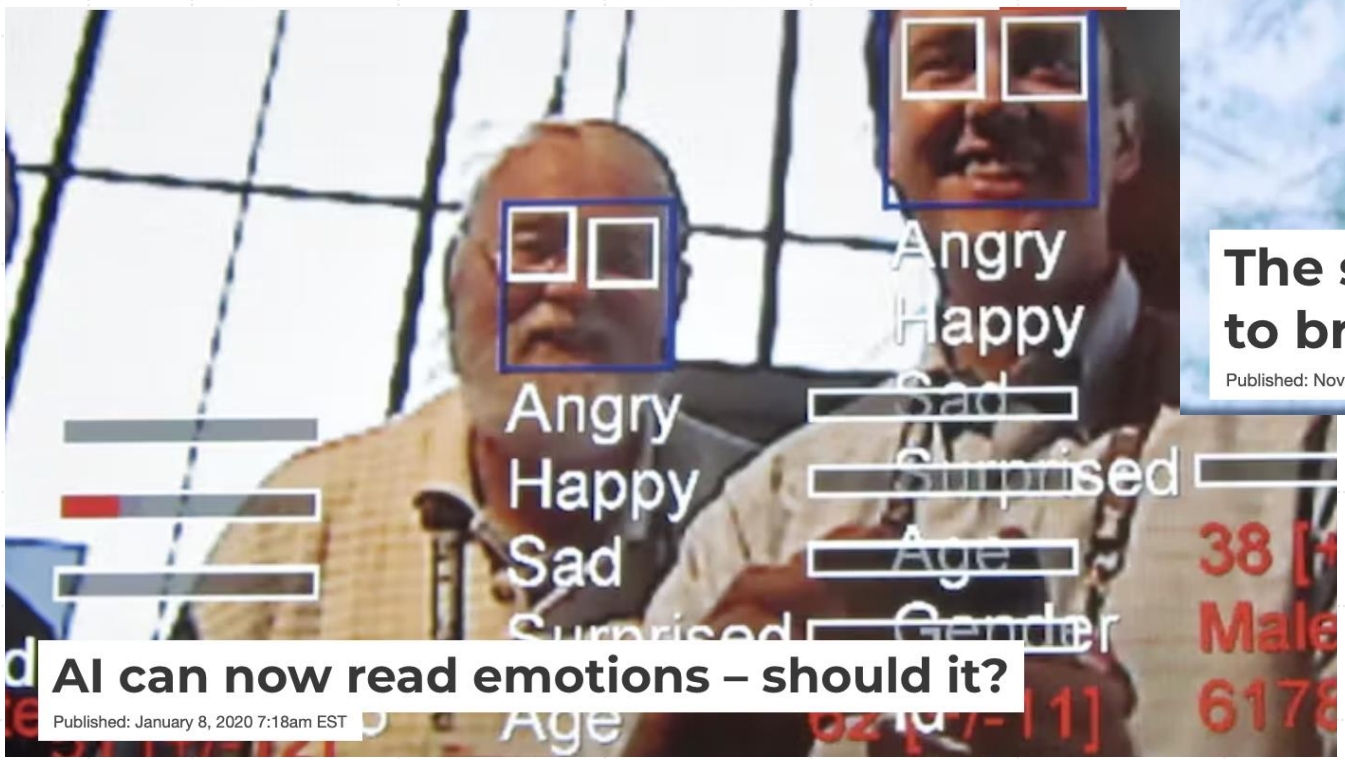
Data is used to educate AI system.

AI system makes informed decisions.



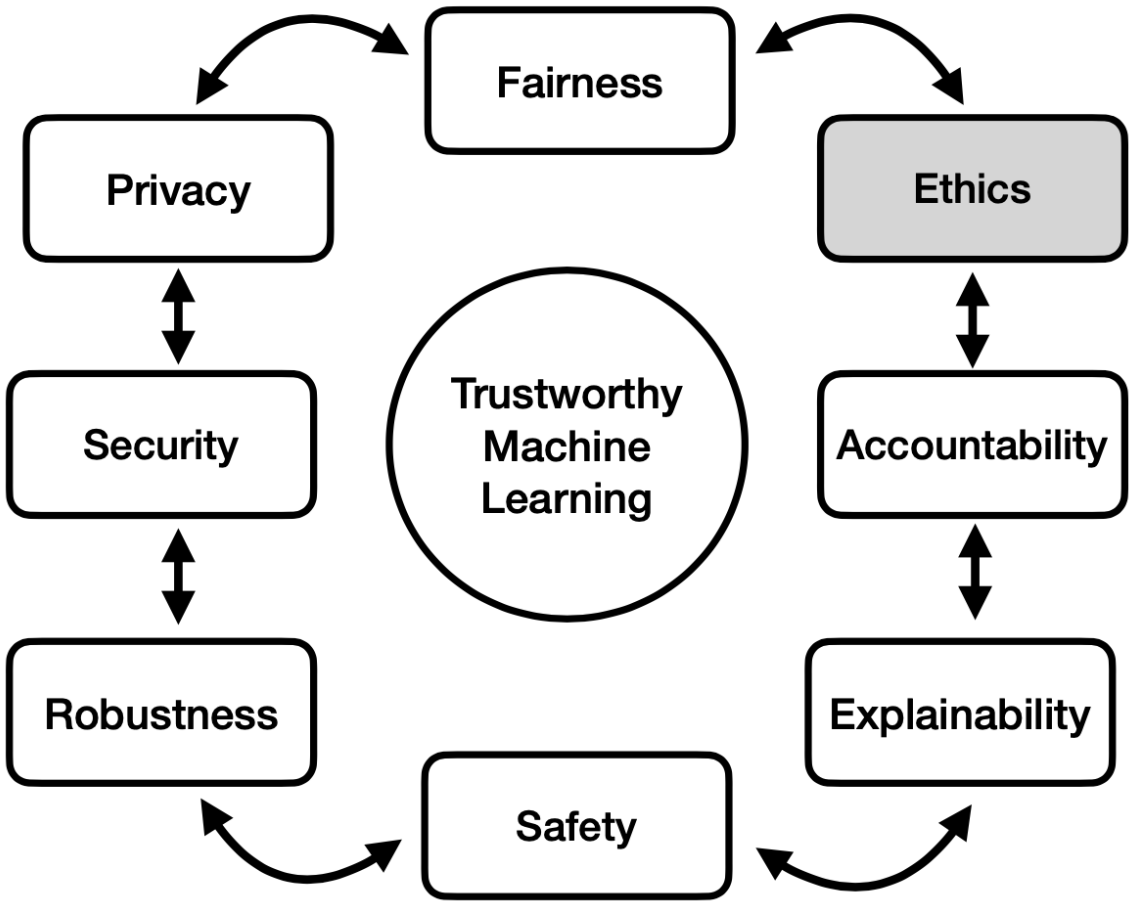
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

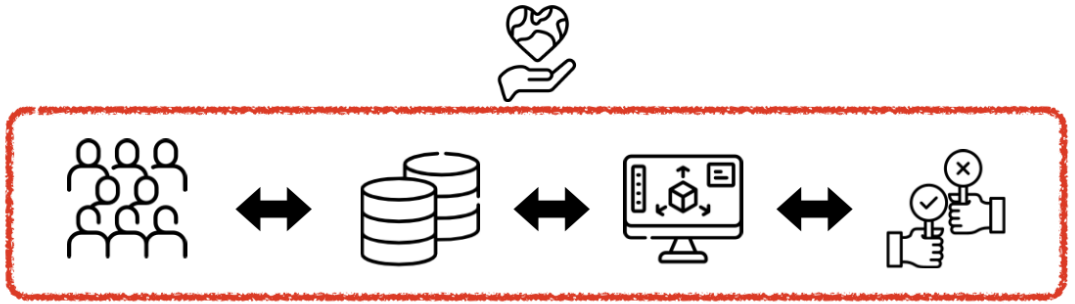
Ethics


Angry
Happy
Sad
Surprised
Age
Gender
Angry
Happy
Sad
Surprised
Age
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

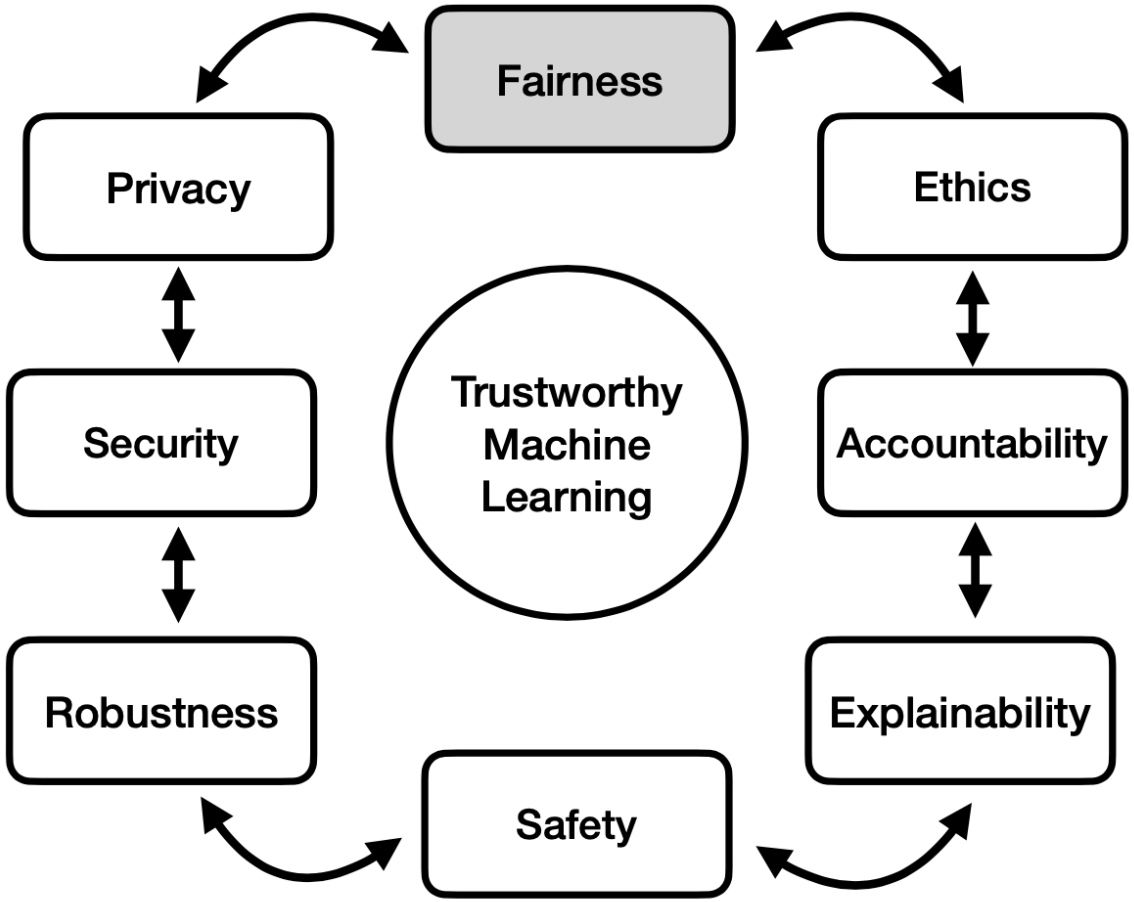
Fairness

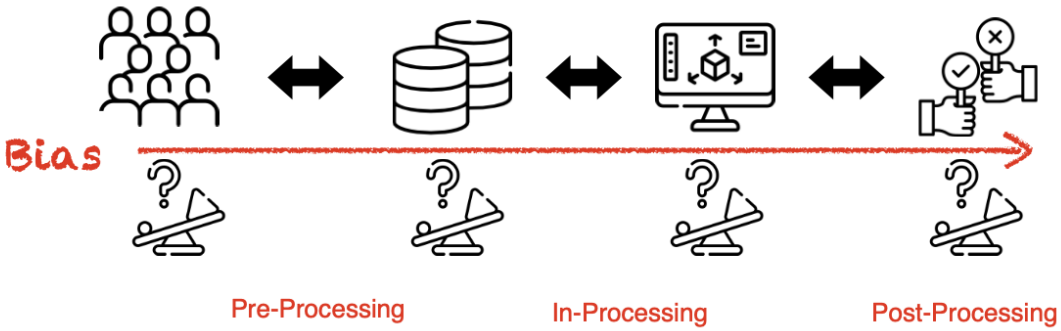
ACLU finds Amazon's facial recognition AI is racially biased

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...

27 July 2018 | Amazon

Fairness





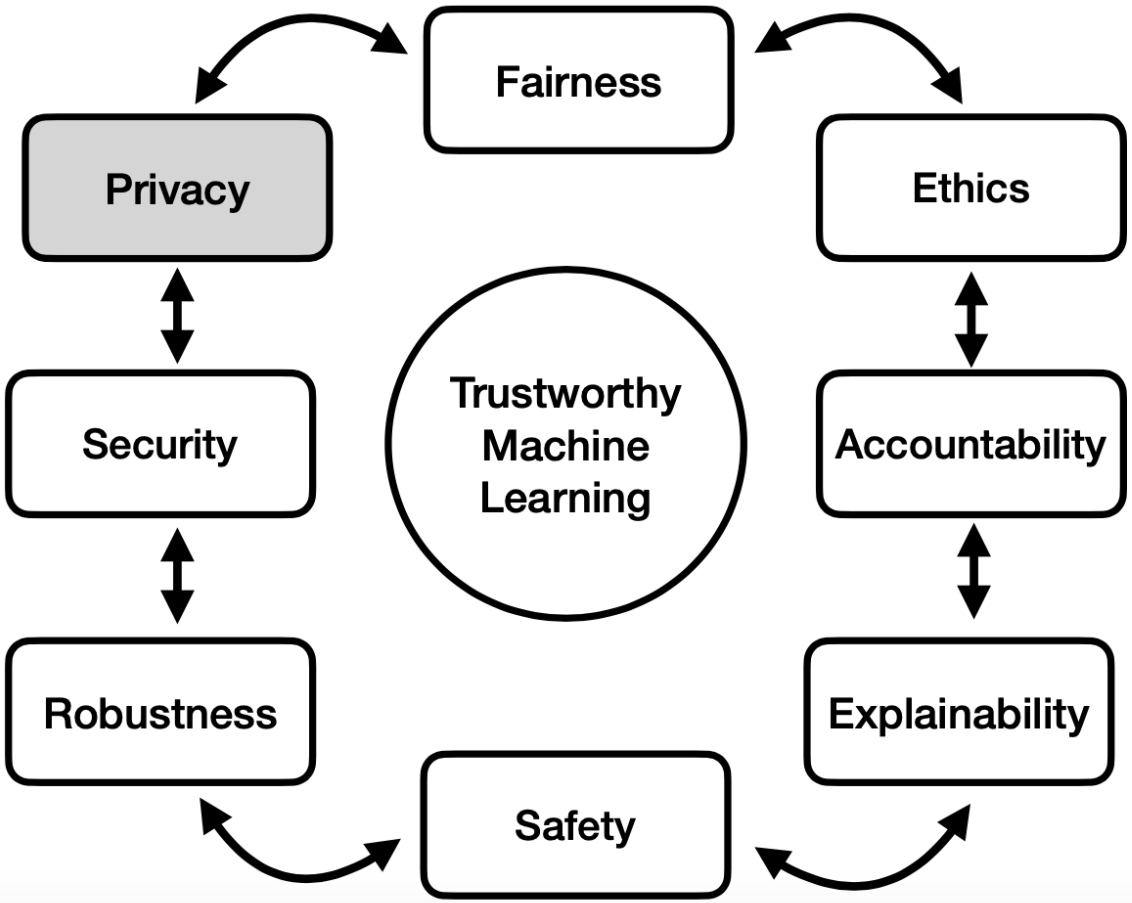
- Fairness in machine learning refers to the various attempts at correcting algorithmic bias in automated decision processes.

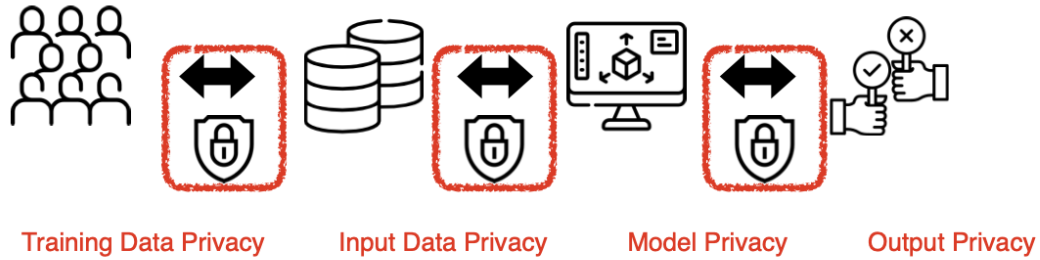
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

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

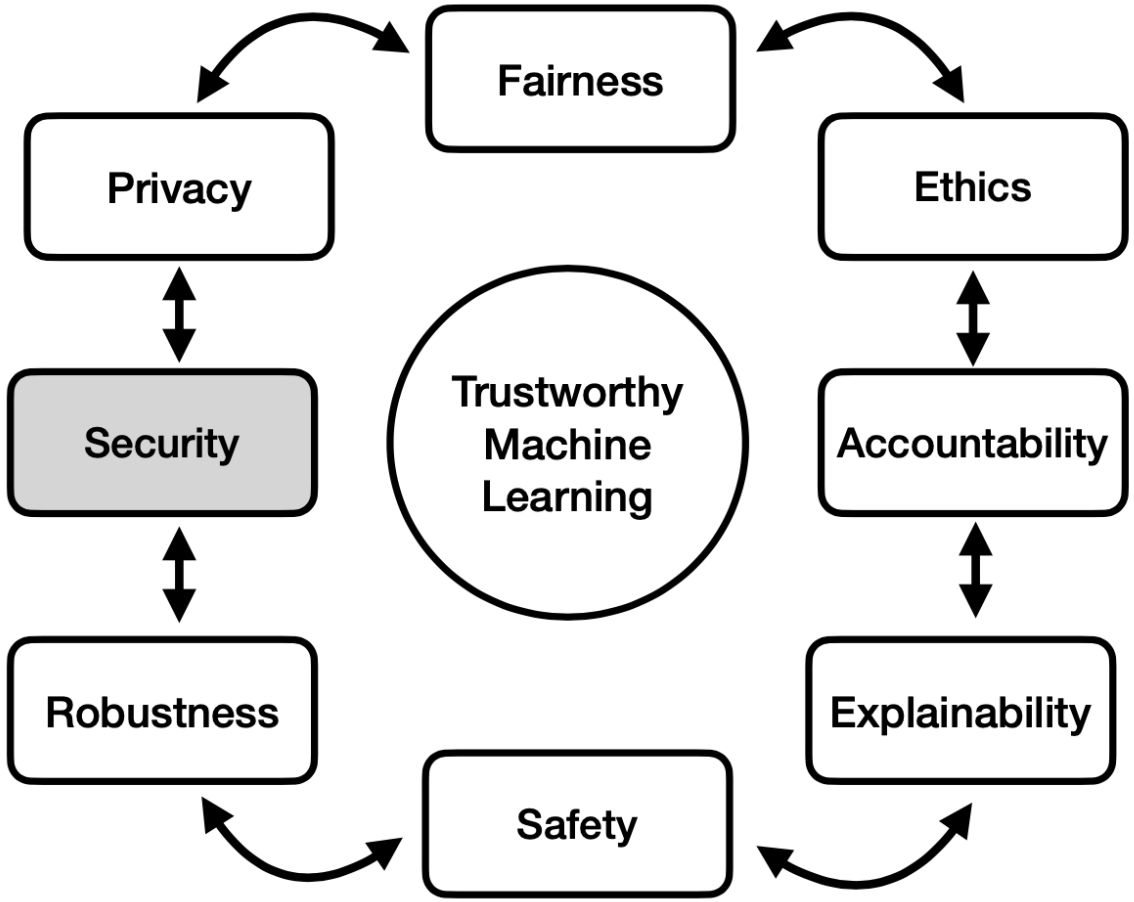
With cyber attacks on a steady incline, something needs to be done so that similar fates can be avoided and patient trust can be prioritized. There are four key best practices that, with the right data access technology in place, can protect healthcare companies from hacks and attacks

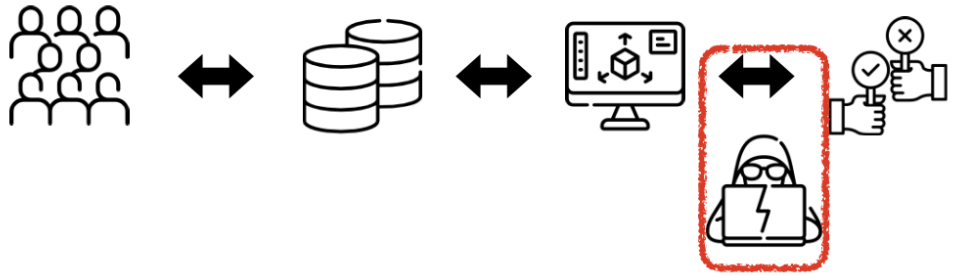
By ELDAD CHAI

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

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

Security





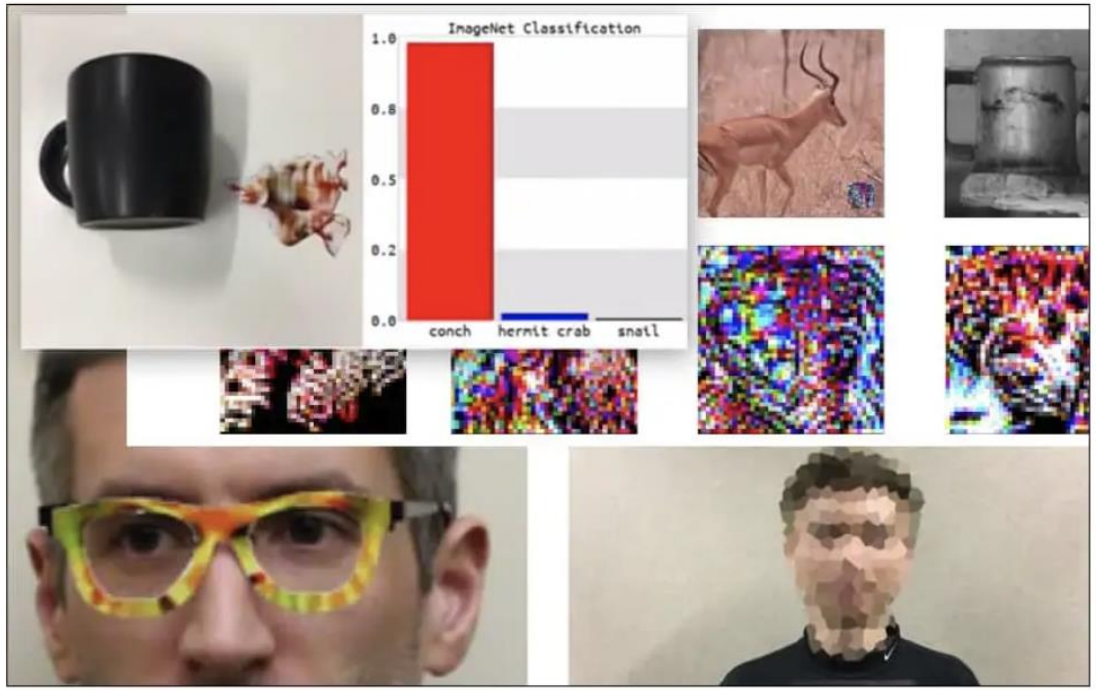
- Cybersecurity is the practice of protecting systems, networks, and programs from digital attacks.

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.

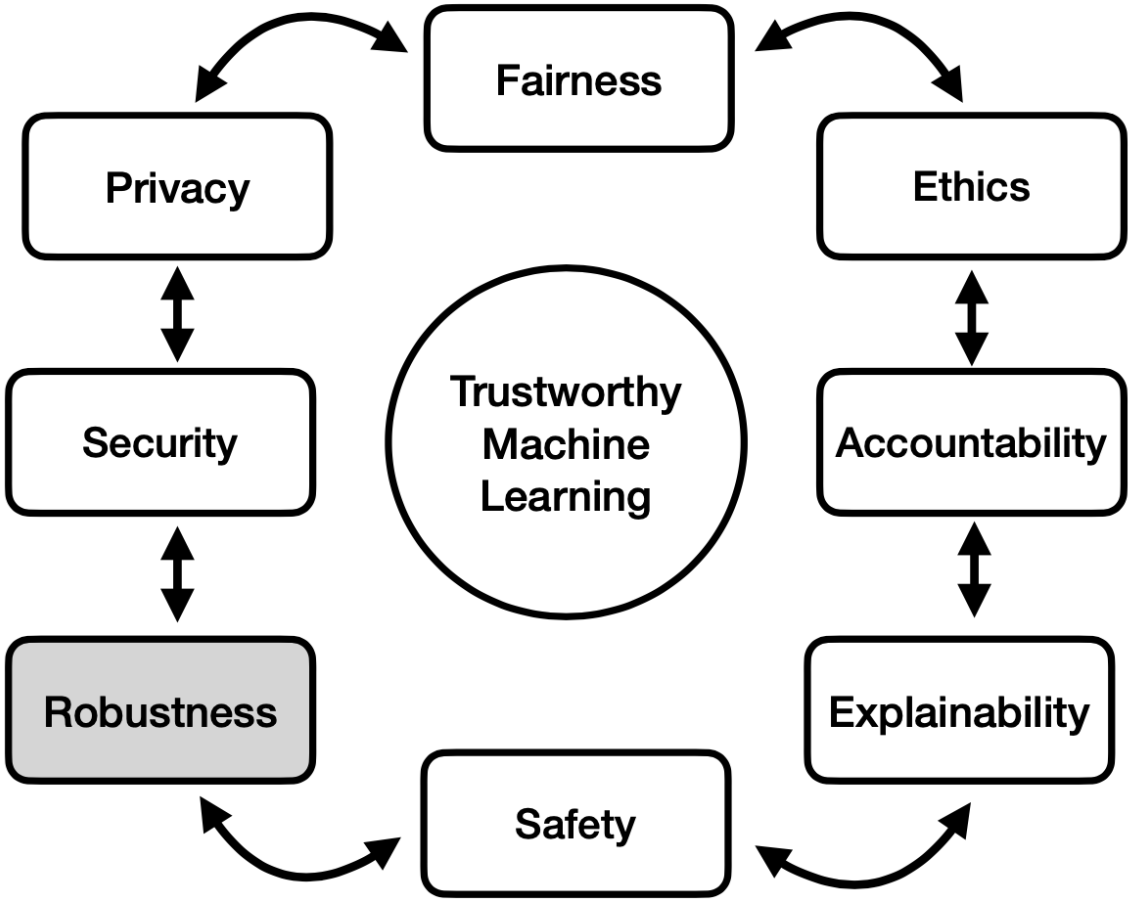
Physical adversarial example from CVPR 2018 paper

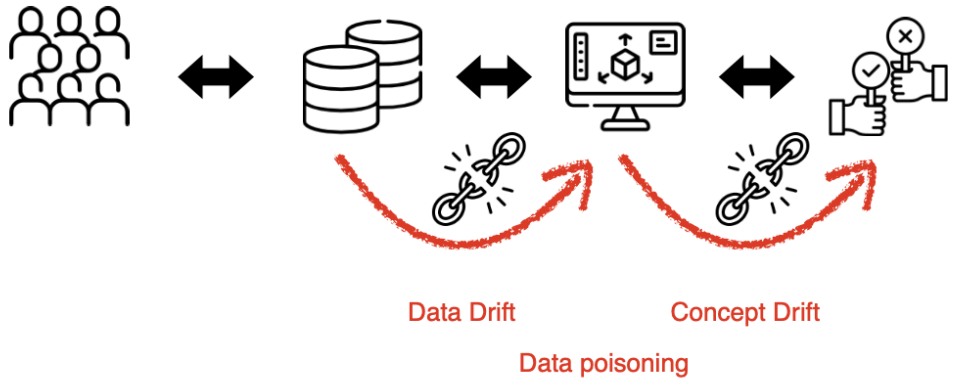
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Can we fool AI?

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

Robustness





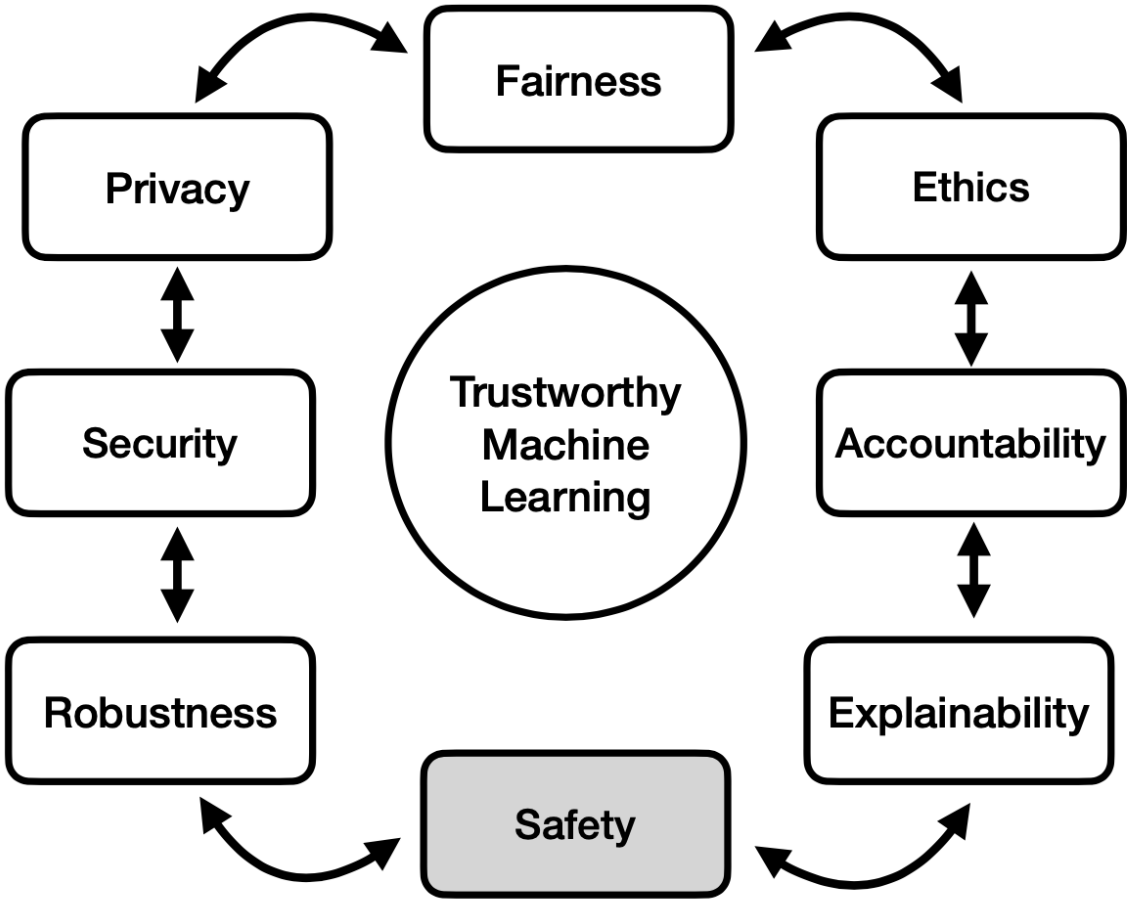
- The robustness is the property that characterizes how effective your algorithm is while being tested on the new independent (but similar) dataset.

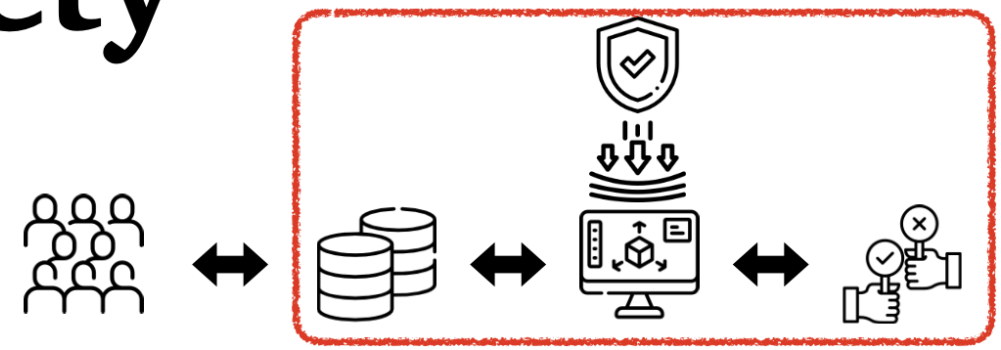
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

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

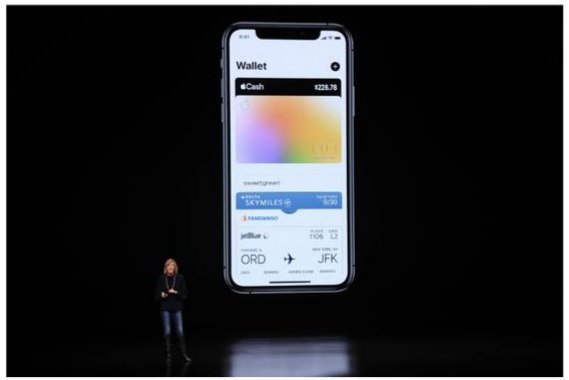
April 28, 2022

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



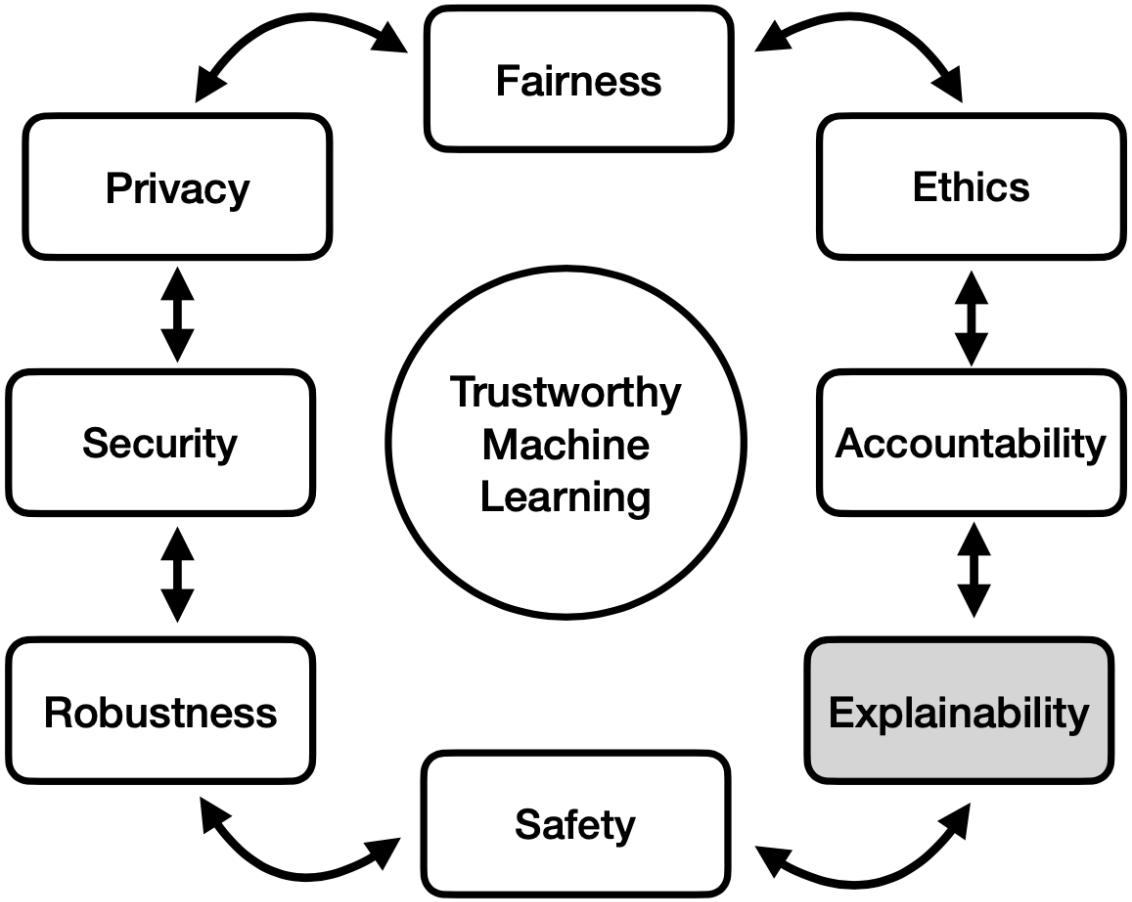
Jennifer Bailey, vice president of Apple Pay. Regulators are investigating Apple Card's algorithm, which is used to determine applicants' creditworthiness. Jim Wilson/The New York Times

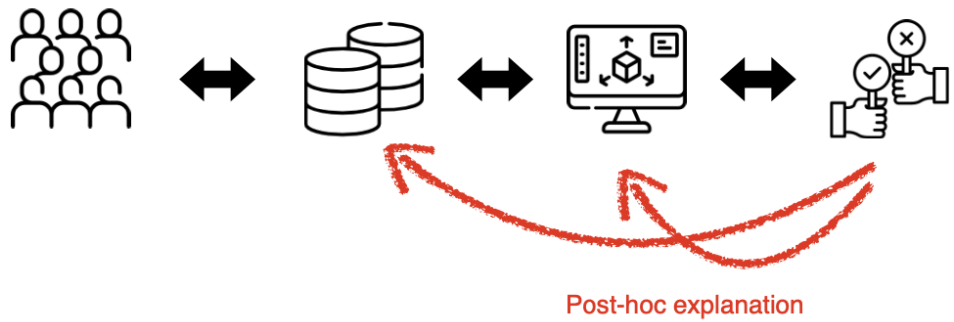
By Neil Vigdor

Nov. 10, 2019

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

Explainability



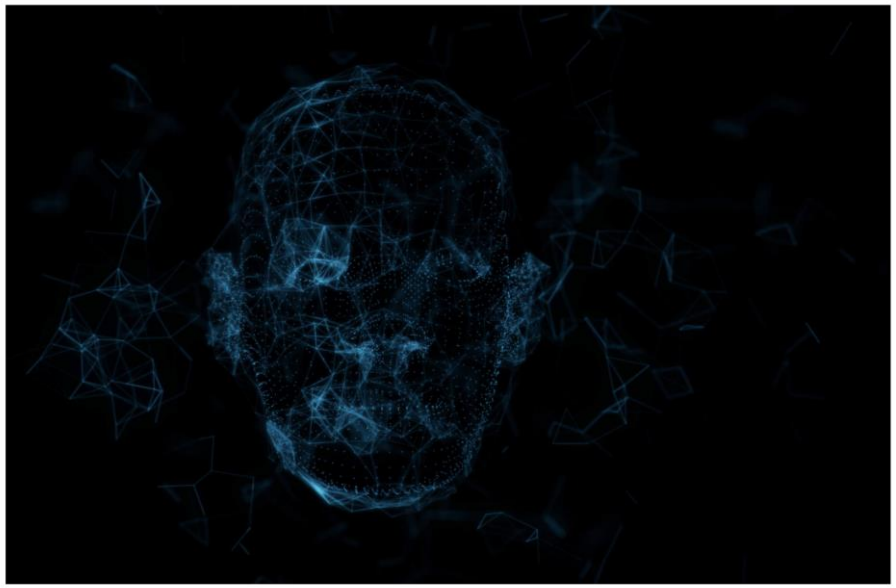


- Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms.

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

Accountability

Error-prone facial recognition leads to another wrongful arrest



About the Author

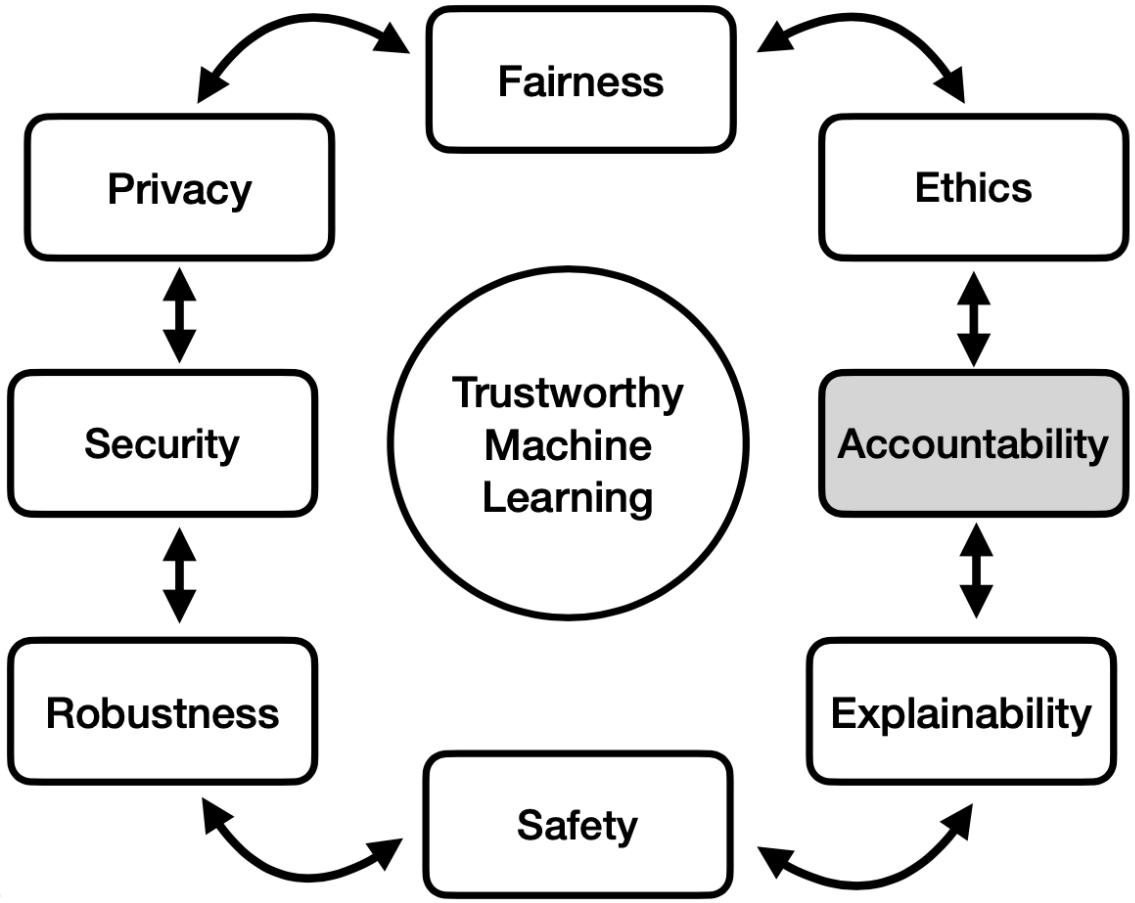
By Ryan Daws | August 7, 2023
Categories: Applications, Artificial Intelligence, Ethics & Society, Face Recognition, Privacy, Surveillance,

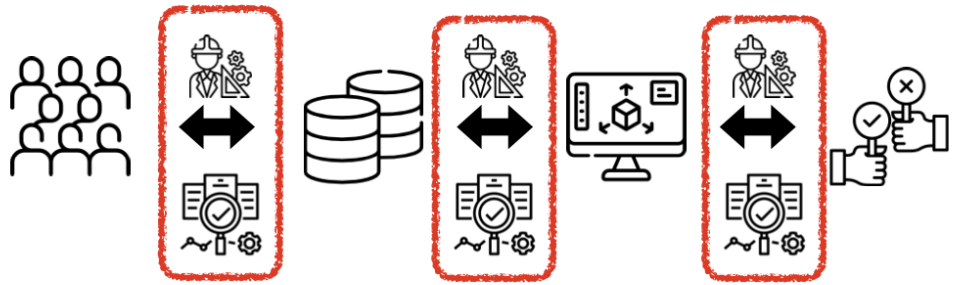
Social media algorithms are still failing to counter misleading content

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

Trustworthy AI: Ethics

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

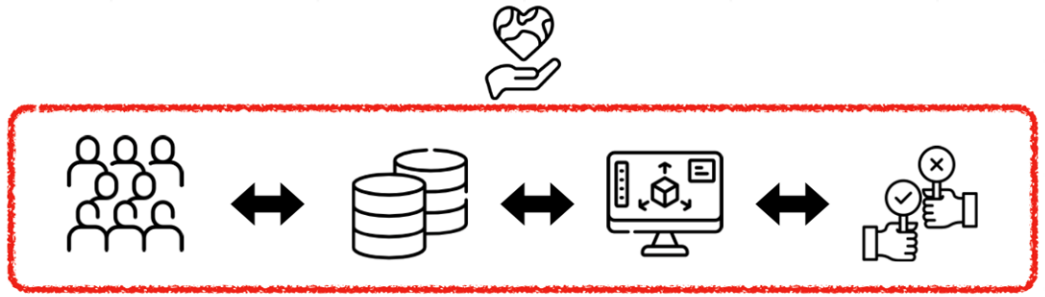
Meharry Medical College

Overview

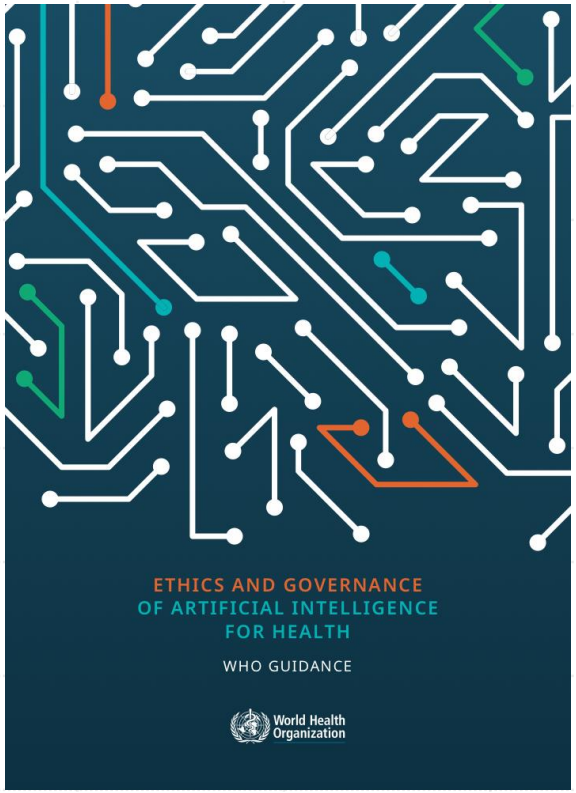
- What is Ethics?
- Establishing Guidelines
- Ethical Principles
  - Transparency
  - Justice & Fairness
  - Non-maleficence
  - Responsibility
  - Privacy

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



Ethical guidelines vary from country to country and across different industries.

Table 2 | Ethics guidelines for AI by country of issuer (USA, international, EU and N/A)

Name of document/website	Issuer	Country of issuer
Unified Ethical Frame for Big Data Analysis. IAF Big Data Ethics Initiative, Part A	The Information Accountability Foundation	USA
The AI Now Report. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term	AI Now Institute	USA
Statement on Algorithmic Transparency and Accountability	Association for Computing Machinery (ACM)	USA
AI Principles	Future of Life Institute	USA
AI—Our Approach	Microsoft	USA
Artificial Intelligence. The Public Policy Opportunity	Intel Corporation	USA
IBM's Principles for Trust and Transparency	IBM	USA
OpenAI Charter	OpenAI	USA
Our Principles	Google	USA
Policy Recommendations on Augmented Intelligence in Health Care H-480.940	American Medical Association (AMA)	USA
Everyday Ethics for Artificial Intelligence. A Practical Guide for Designers and Developers	IBM	USA
Governing Artificial Intelligence. Upholding Human Rights & Dignity	Data & Society	USA
Intel's AI Privacy Policy White Paper. Protecting Individuals' Privacy and Data in the Artificial Intelligence World	Intel Corporation	USA
Introducing Unity's Guiding Principles for Ethical AI—Unity Blog	Unity Technologies	USA
Digital Decisions	Center for Democracy & Technology	USA
Science, Law and Society (SLS) Initiative	The Future Society	USA
AI Now 2018 Report	AI Now Institute	USA
Responsible Bots: 10 Guidelines for Developers of Conversational AI	Microsoft	USA
Preparing for the Future of Artificial Intelligence	Executive Office of the President; National Science and Technology Council; Committee on Technology	USA
The National Artificial Intelligence Research and Development Strategic Plan	National Science and Technology Council; Networking and Information Technology Research and Development Subcommittee	USA
AI Now 2017 Report	AI Now Institute	USA

Image from Jobin et al. (2019). *The global landscape of AI ethics guidelines*.

Overarching Ethical Principles

Transparency

Justice and Fairness

Non-maleficence

Responsibility

Privacy

Transparency

AI should be interpretable for users.

Explainability describes the extent to which human users can comprehend AI systems and trust the outcomes.

Justice and Fairness

AI should avoid unwanted bias and discrimination.

Algorithmic discrimination indicates the presence of biases.
Fairness refers to the attempts used to correct algorithmic biases.

Non - maleficence

AI should never cause harm.

AI safety identifies causes of unintended behavior and potential harm and develops tools to ensure systems work safely.

Responsibility

AI should make reliable decisions that are accountable.

Accountability refers to the ability to ascertain whether an AI system is behaving properly.

Privacy

The data obtained by AI should be secure and protected.

Privacy-preserving machine learning utilizes techniques to safeguard data to help prevent the exposure of sensitive data.

Thank You

Please send us your questions at:

vgupta@mmc.edu and

dpounds24@email.mmc.edu

Trustworthy AI: Fairness

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

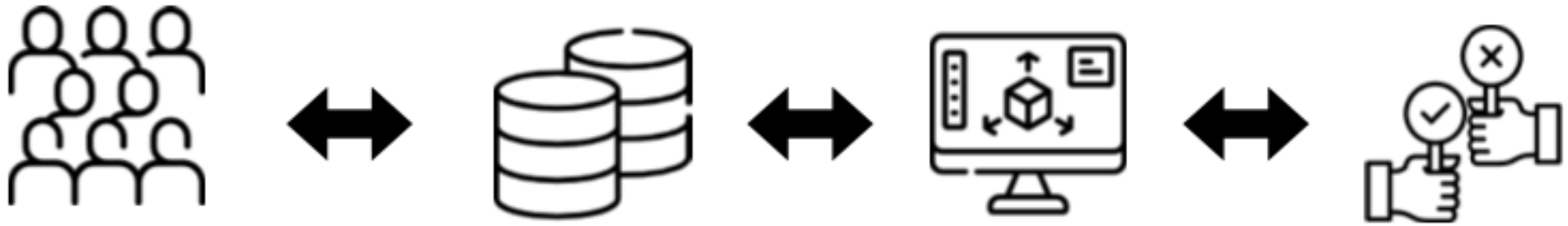
Meharry Medical College

Overview

- Algorithmic Bias/Discrimination
- Categories of Bias
  - Data Bias
  - Interpretation Bias
- Examples of Bias
- Addressing Fairness in ML

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

Algorithmic discrimination is the result of bias.



bias

Biases introduced during the machine learning process come in different forms.

Data Bias

Algorithms are trained using biased data.

Examples:

- Selection bias
- Sampling bias
- Reporting bias
- Participation bias
- Non-response bias
- Coverage bias

Interpretation Bias

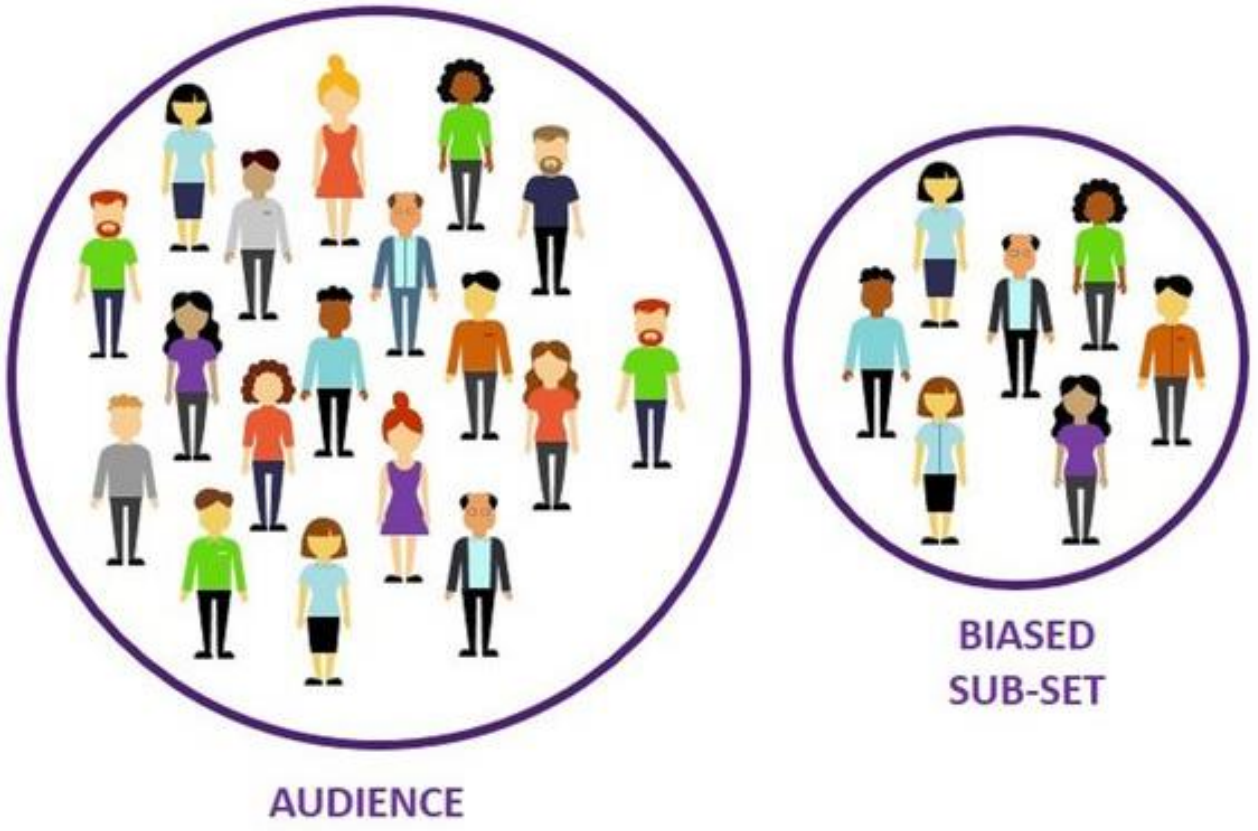
Human assumptions perpetuate a skewed interpretation of results.

Examples:

- Automation bias
- Overgeneralization
- Confirmation bias
- Experimenter's bias
- Group attribution bias
- Implicit bias
- Correlation fallacy

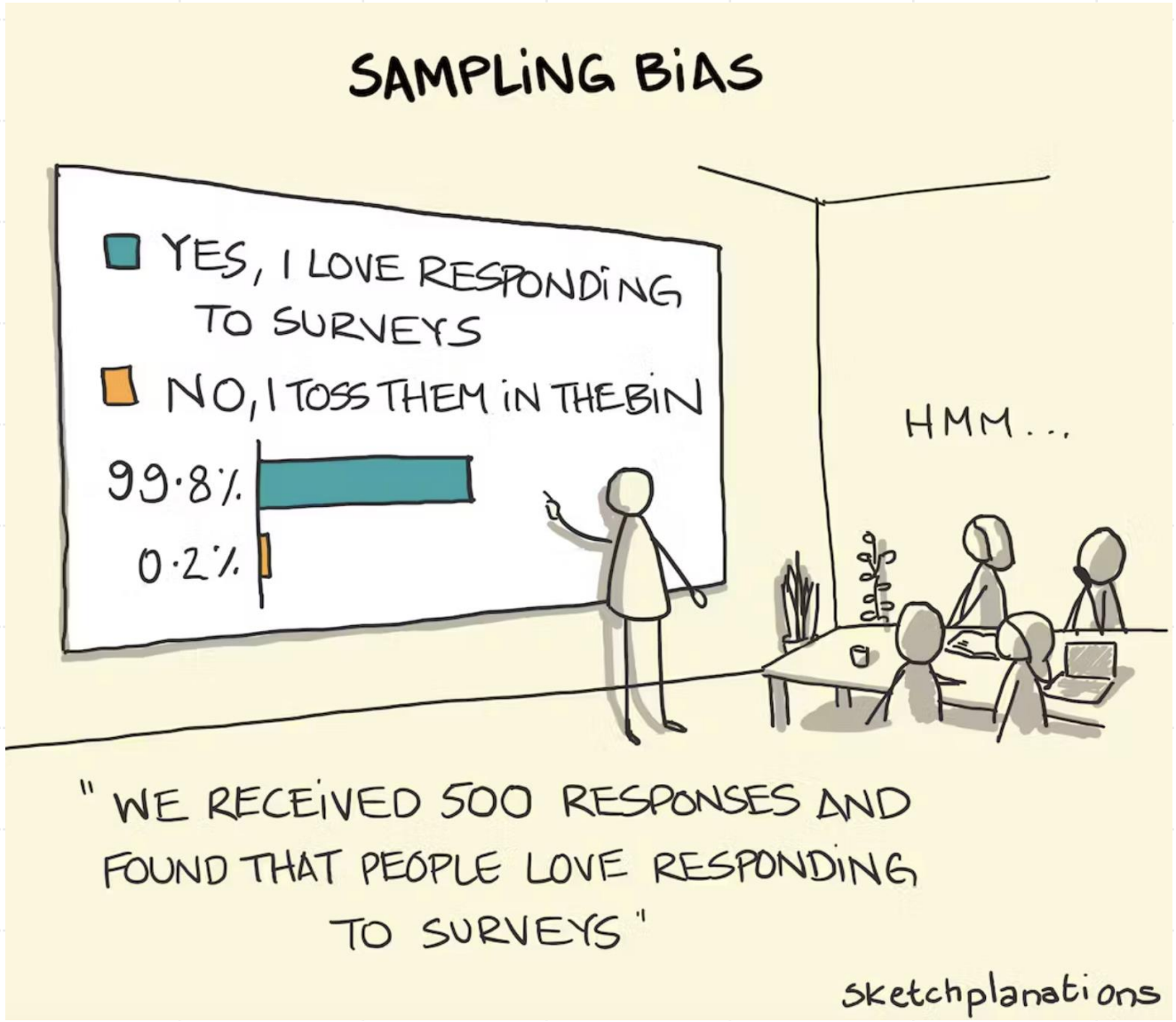
Data Bias

- Selection Bias – Data selections don’t reflect randomization



Data Bias

- Sampling bias – data instances are more frequently samples

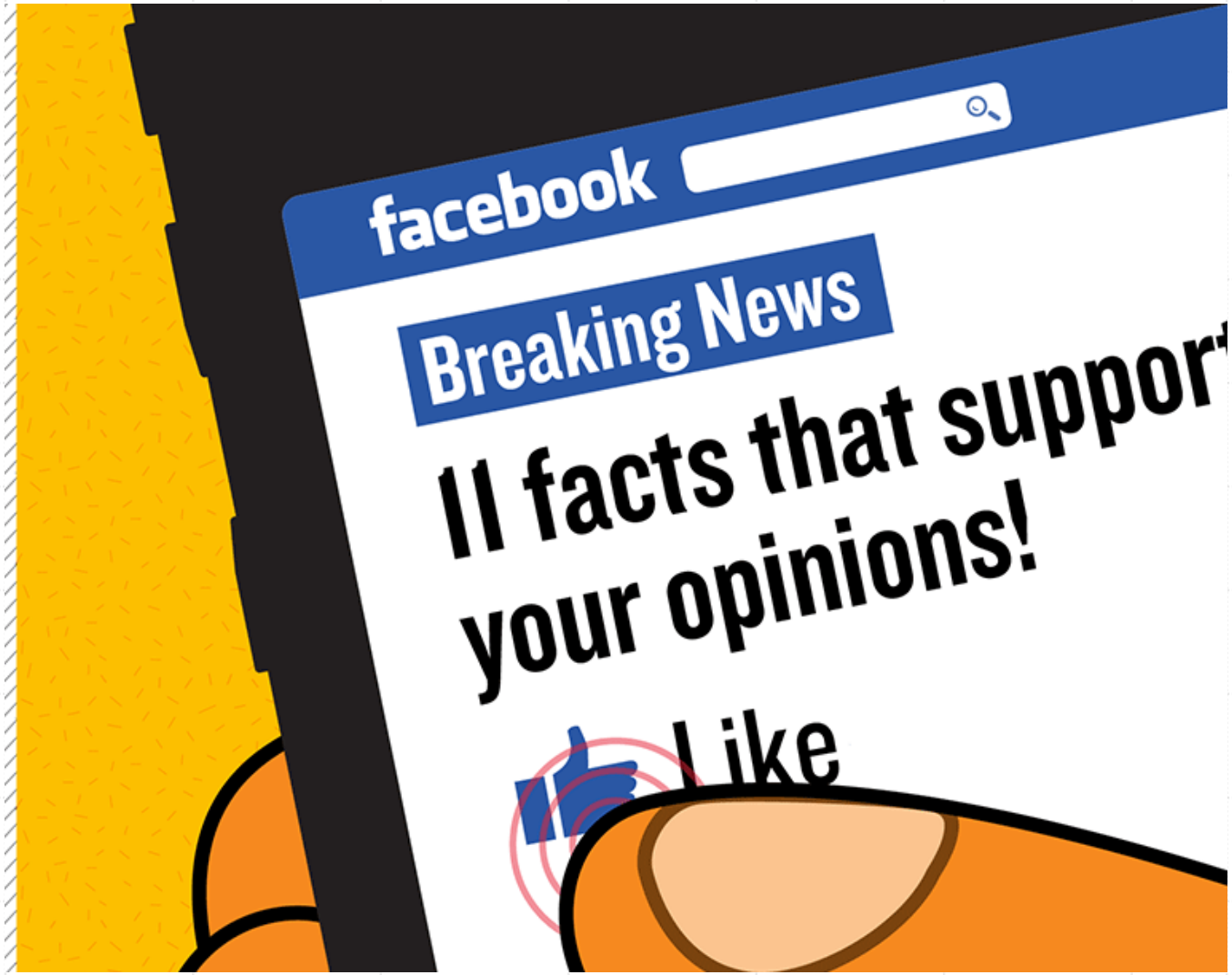


Interpretation Bias

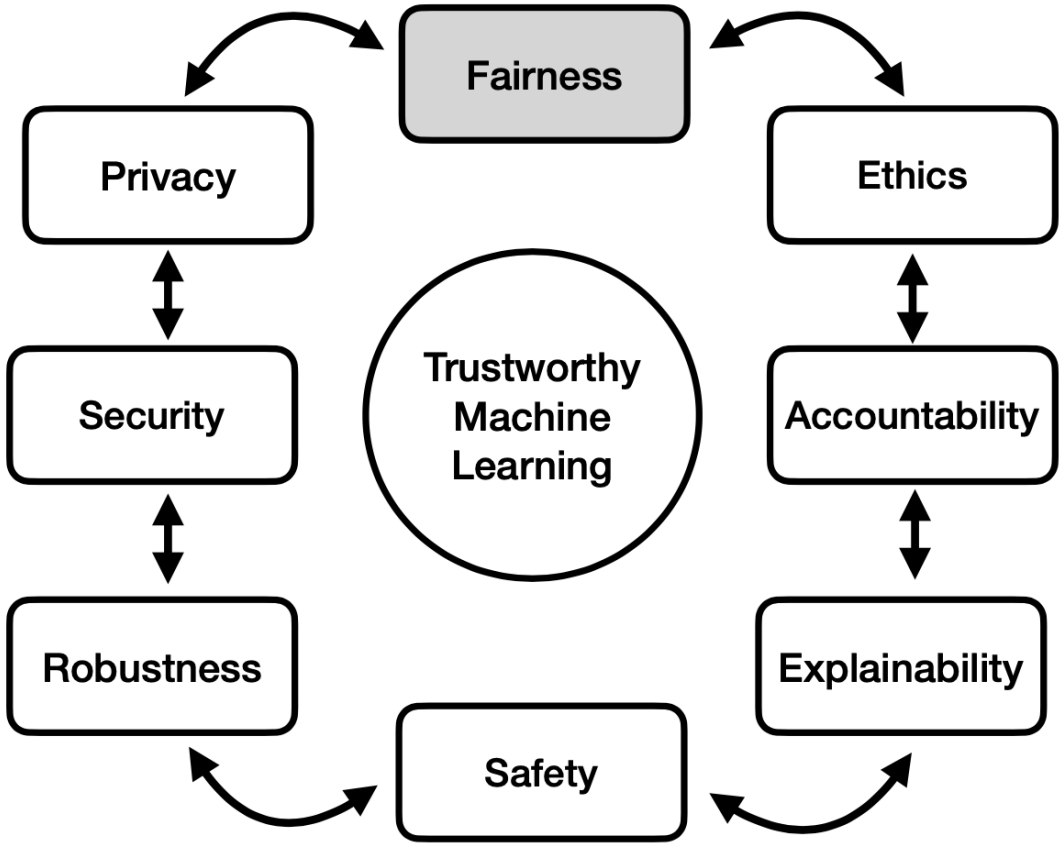
- Overgeneralization – making more general conclusions from limited testing data

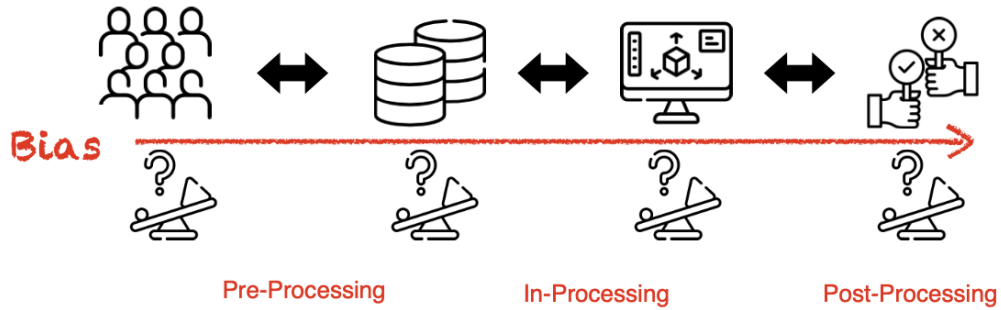
Interpretation bias

- Confirmation bias – the tendency to search for, interpret, favor, recall information in a way that confirms pre-existing beliefs



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

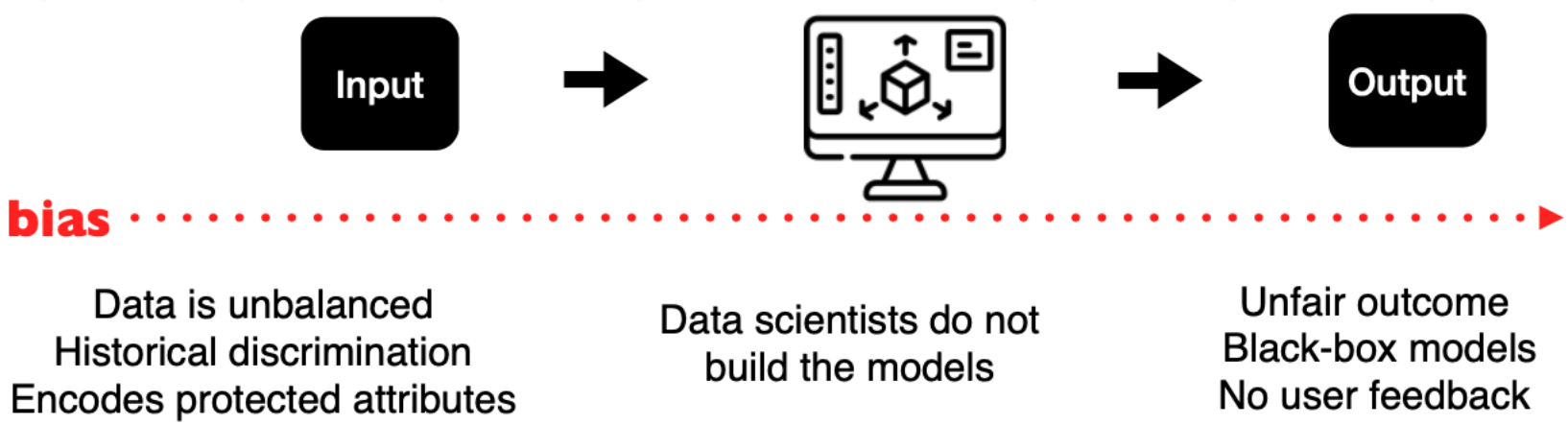
Fairness is not a simple concept.

Correcting bias requires:

- Identifying the type of bias present
- Deciding what properties would indicate an unbiased process

Biases are subjective, therefore, there is no single agreed-upon measure of fairness.

Addressing Fairness in Machine Learning



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

Addressing Fairness in Machine Learning

Data is unbalanced
Historical discrimination
Encodes protected attributes

Data scientists do not build the models

Unfair outcome
Black-box models
No user feedback

Pre-processing

Discrimination Discovery
Sampling/weighting
Representation learning

In-processing

Learning/Inference subject to constraints
Ranking
Constrained Optimization

Post-processing

Thresholding
Causal discovery
Explainable AI
Verification

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

Trustworthy AI: Privacy

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

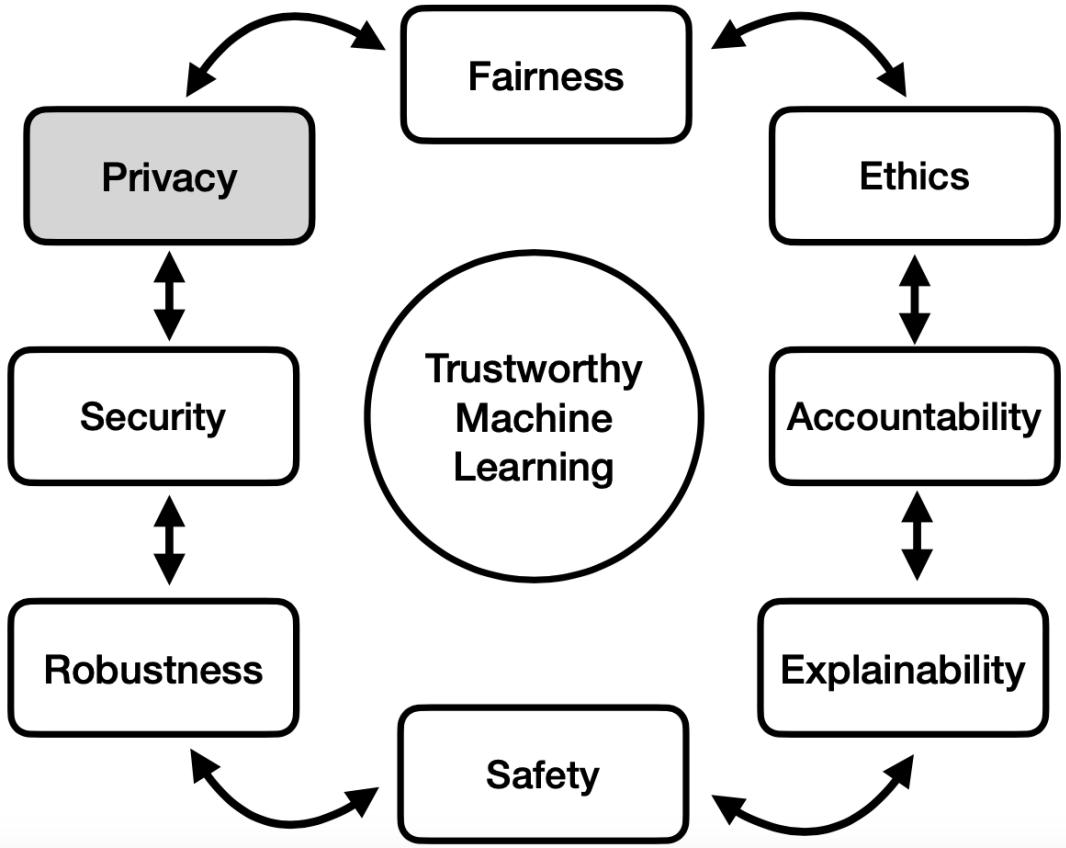
Overview

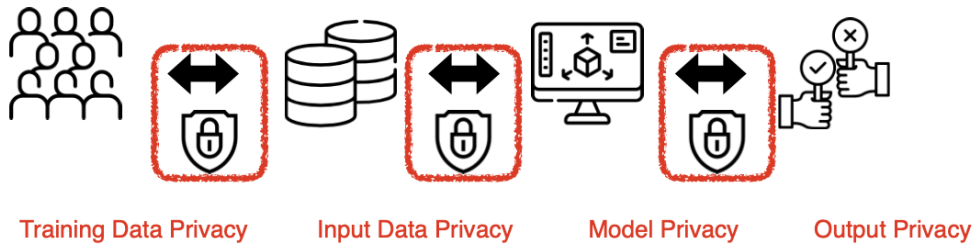
- Defining Privacy
- Privacy-Preserving Machine Learning
- PPML Techniques
  - Data Anonymization
  - Differential Privacy
  - Homomorphic Encryption
  - Secure Multi-party Computation
  - Federated Learning

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

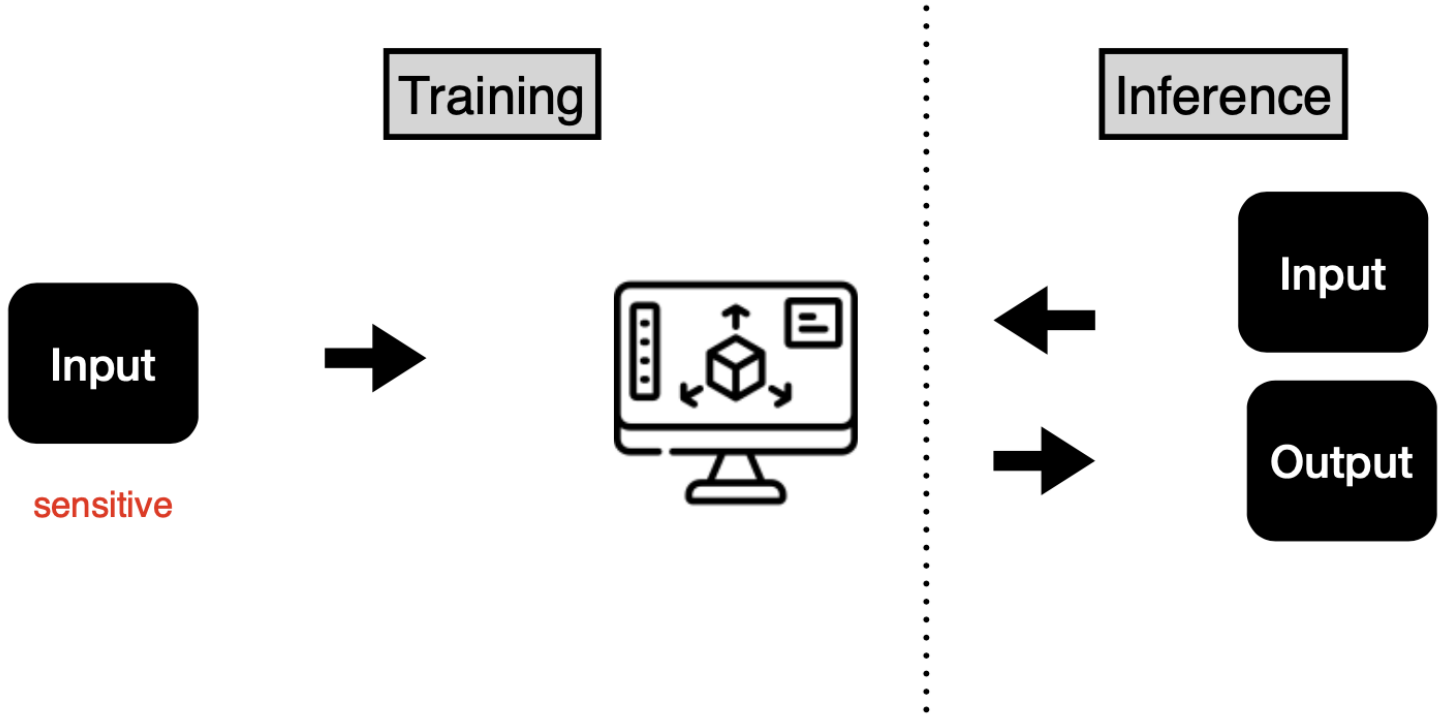




- Data privacy is a central issue to training and testing AI models, especially ones that train and infer on sensitive data.

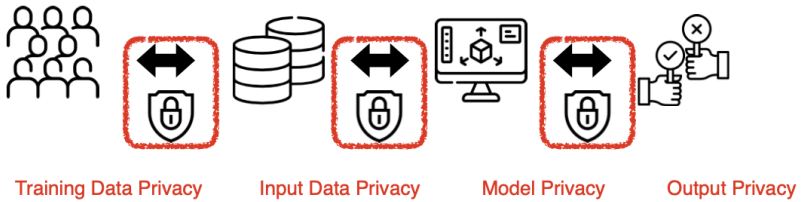
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Privacy Setting in AI/ML



We can gain useful insight about the population without knowing about individuals.

Privacy - Preserving Machine Learning



- Privacy-preserving machine learning (PPML) is a set of techniques and practices that safeguard sensitive data during training and deployment of AI models

Privacy-Preserving Techniques

DATA ANONYMIZATION

DIFFERENTIAL PRIVACY

HOMOMORPHIC ENCRYPTION

SECURE MULTI-PARTY COMPUTATION

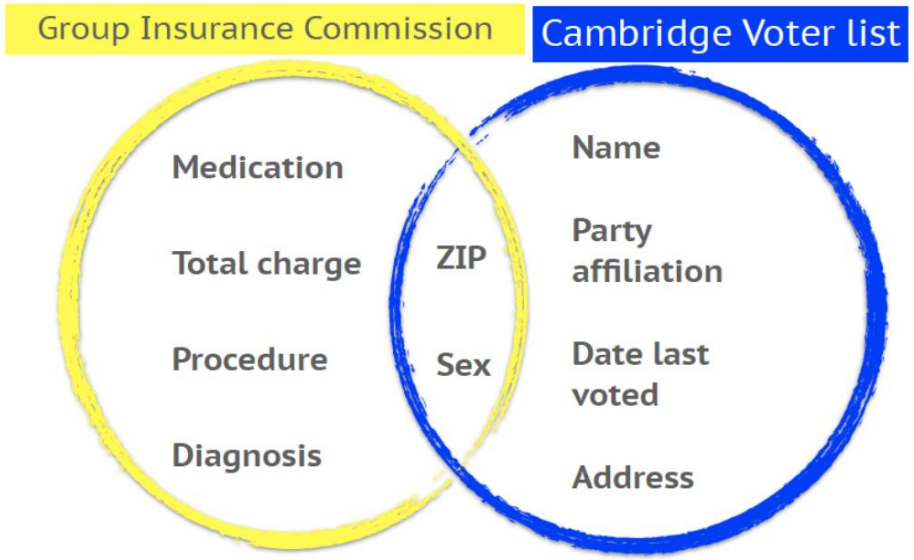
FEDERATED LEARNING

Data Anonymization

Data anonymization is a process of modifying data in a way that eliminates connections to specific individuals.

Name	Sex	Party	Date Last Voted	Address	Zip
Gov. Sam Thomas	M	R	…	…	12345
Lt. Gov. Angie Stevenson	F	R	…	…	12354
Sen. Paul Childs	M	D	…	…	12346
Sen. Lisa Wells	F	D	…	…	12345
Cong. Tim Allen	M	R	…	…	12355
Cong. Rose Smith	F	D	….	…	12345

[Sweeny 02]



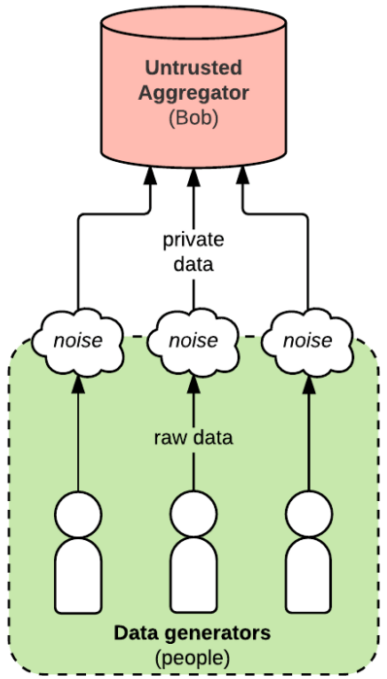
Ex. A male living in zip code 12345 was diagnosed with lung cancer. Who could it be?

Of the six people listed, three are men but only one lives within that zip code.

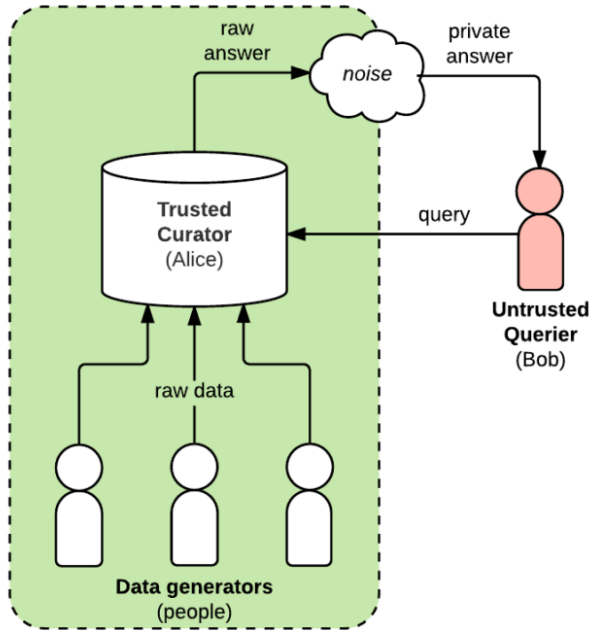
It must be Governor Sam Thomas!

Differential Privacy

- Differential privacy is a framework designed to ensure the privacy of individuals in a dataset. Noise is added to the dataset which makes it difficult for attackers to discern information that is specific to any individual.

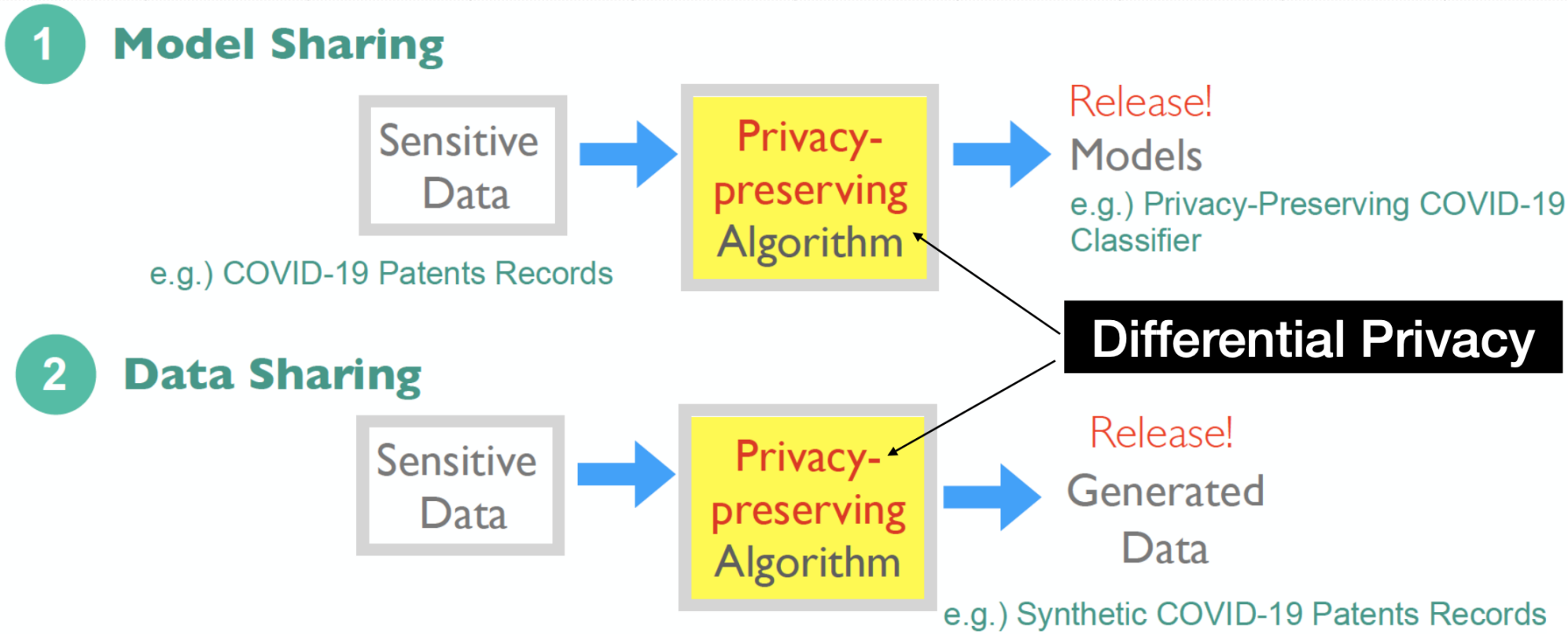


Local privacy



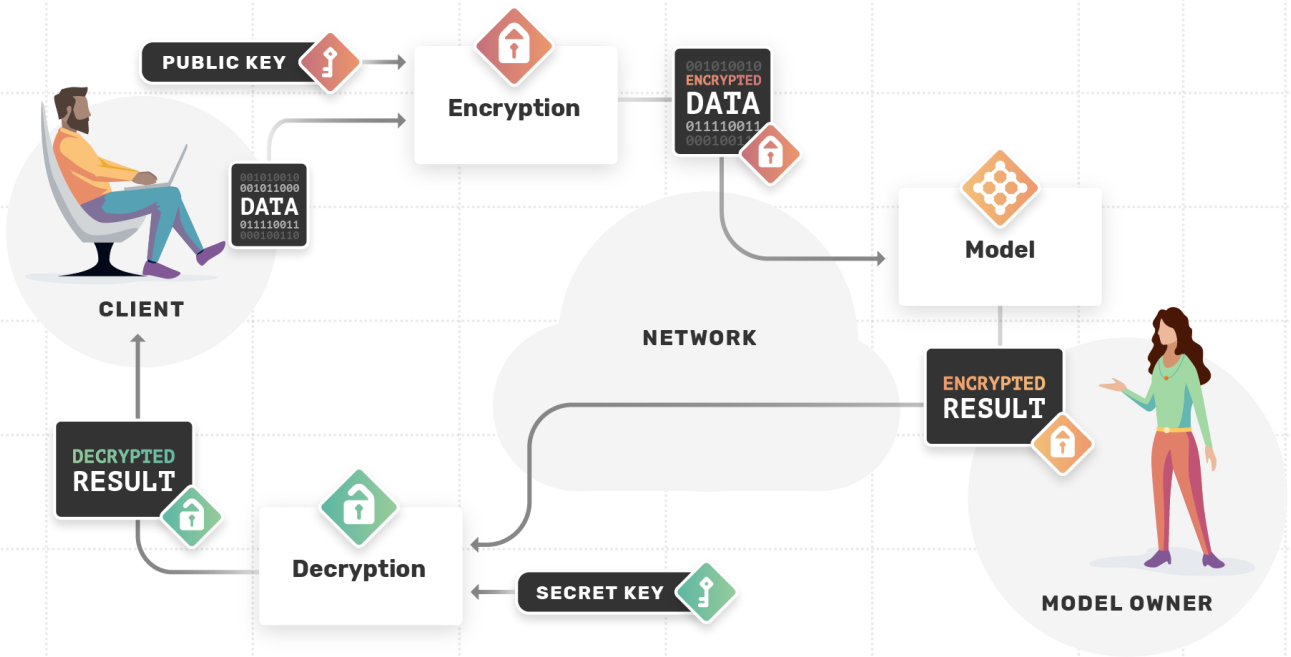
Global privacy

Privacy Settings: Single Data Source

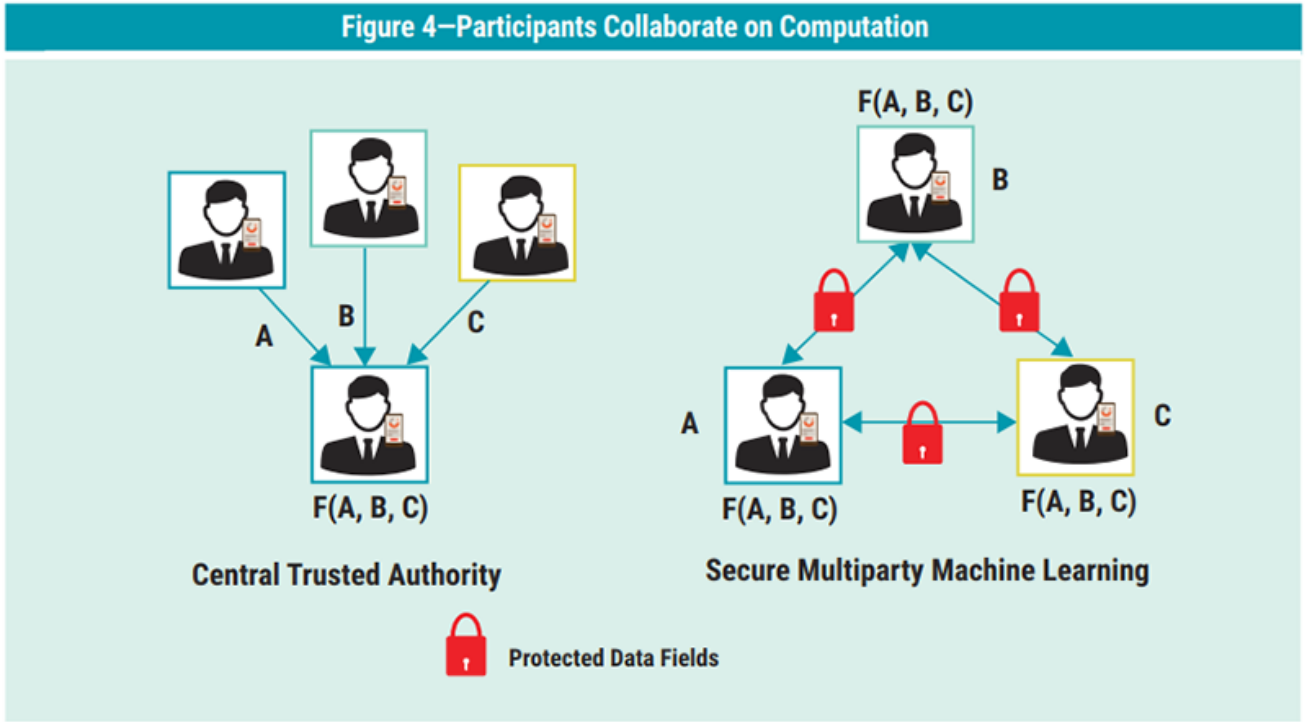


Homomorphic Encryption

- Homomorphic encryption is conversion of data into a coded format that still allows it to be manipulated like original data without compromising encryption.

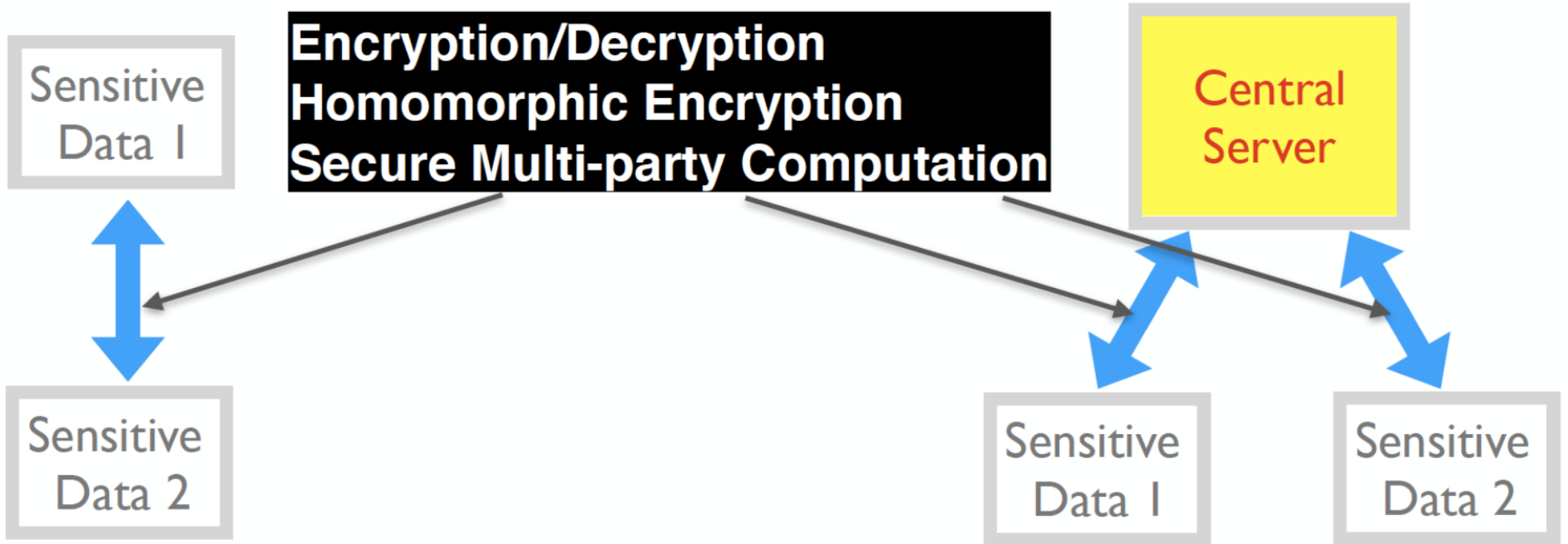


Secure Multi-party Computation



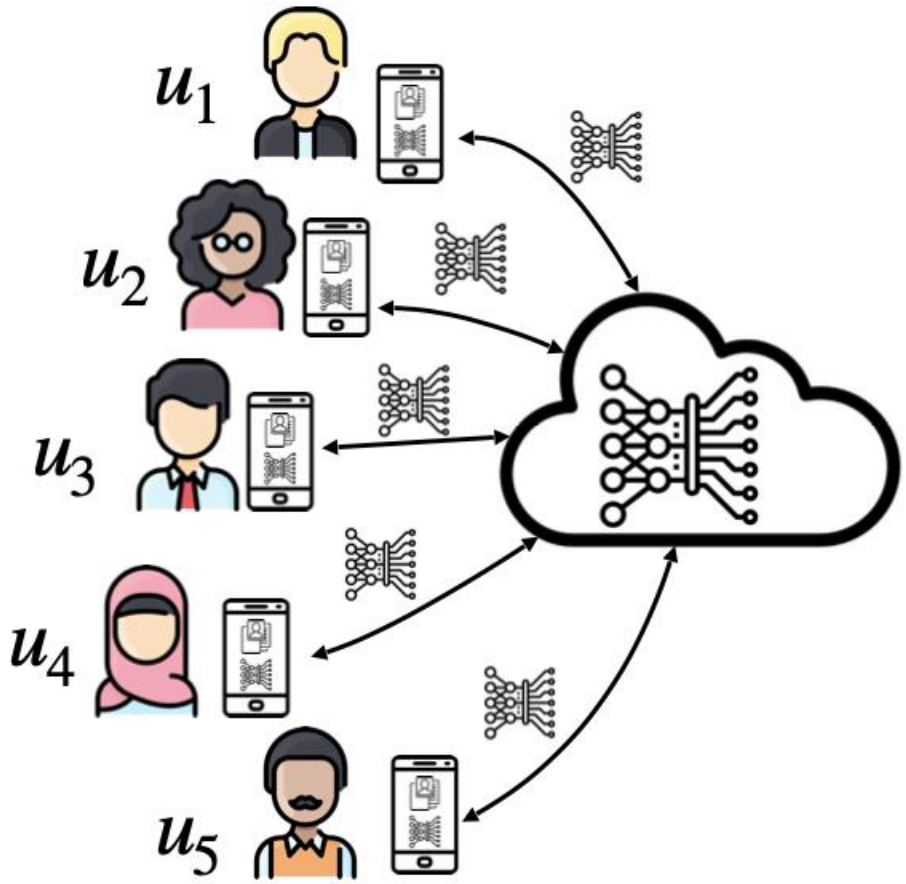
- Secure multi-party computations is a technique that allows multiple parties to collectively perform computations on their combined data while ensuring the privacy of individual information.

Privacy Settings: Multiple Data Sources



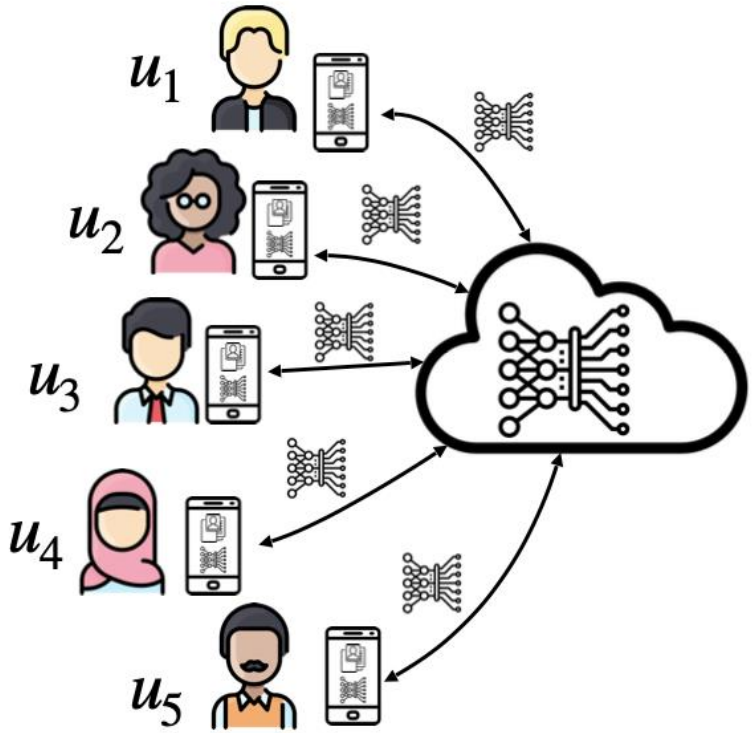
Federated Learning

- Federated learning is a machine learning setting where multiple entities (clients) collaborate in solving a problem, under the coordination of a central server or service provider.



Federated Learning

Data is generated locally and remains decentralized. Each client stores its own data and cannot read the data of other clients.



A central server/service coordinates training, but never sees raw data.

Thank You

Please send us your questions at:

vgupta@mmc.edu and

dpounds24@email.mmc.edu

Trustworthy AI: Security

Destiny Pounds

M.S. Student, Biomedical Data Science

Advised by Vibhuti Gupta, Ph.D.

Assistant Professor, Computer Science and Data Science

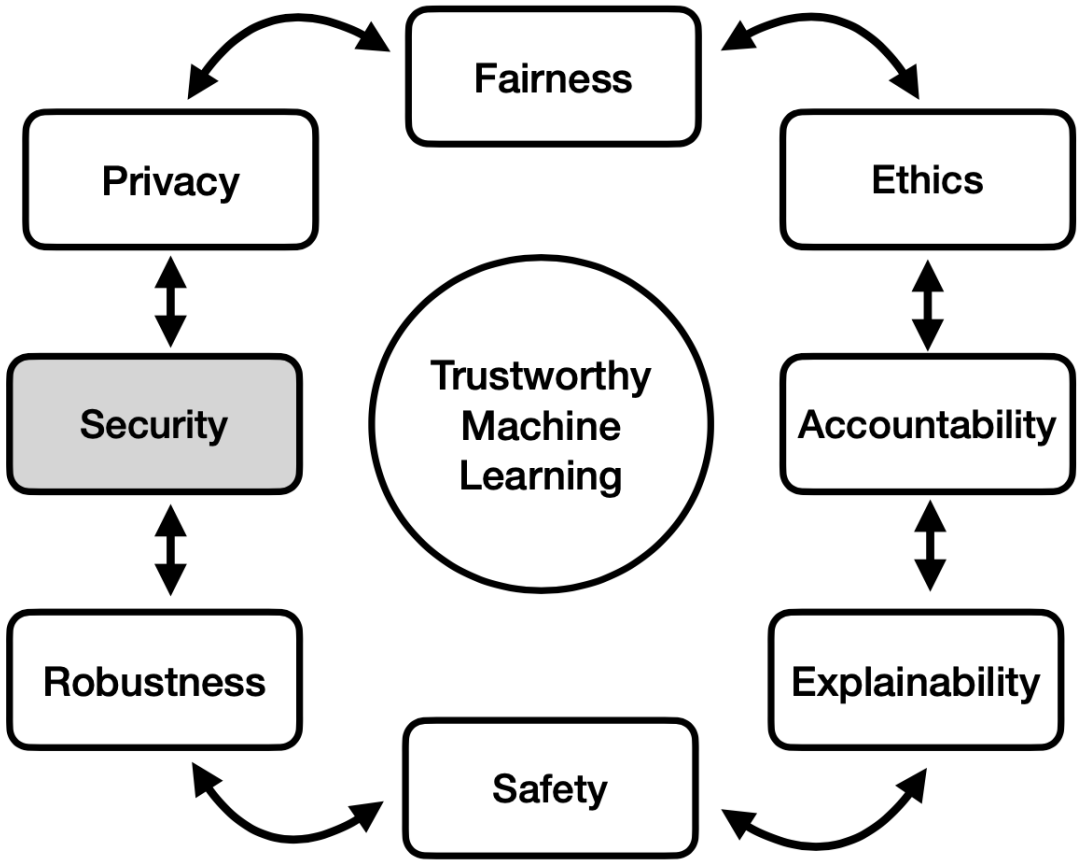
School of Applied Computational Sciences

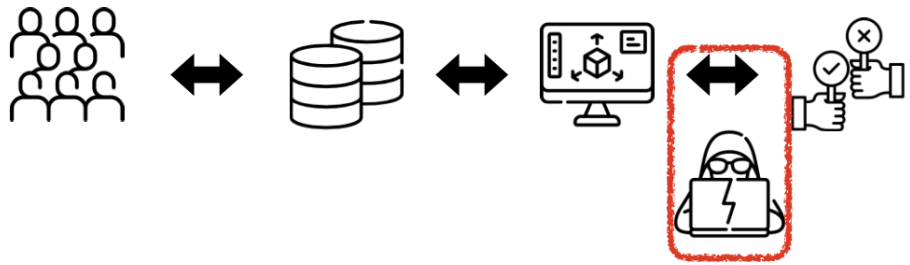
Meharry Medical College

Overview

- Defining Security
- Attack Types
  - Data Poisoning
  - Evasion Attacks
  - Membership Inference Attacks
  - Model Inversion Attacks
  - Model Extraction Attacks
- Mitigation Techniques

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

Attack Types

Data Poisoning

Evasion Attacks

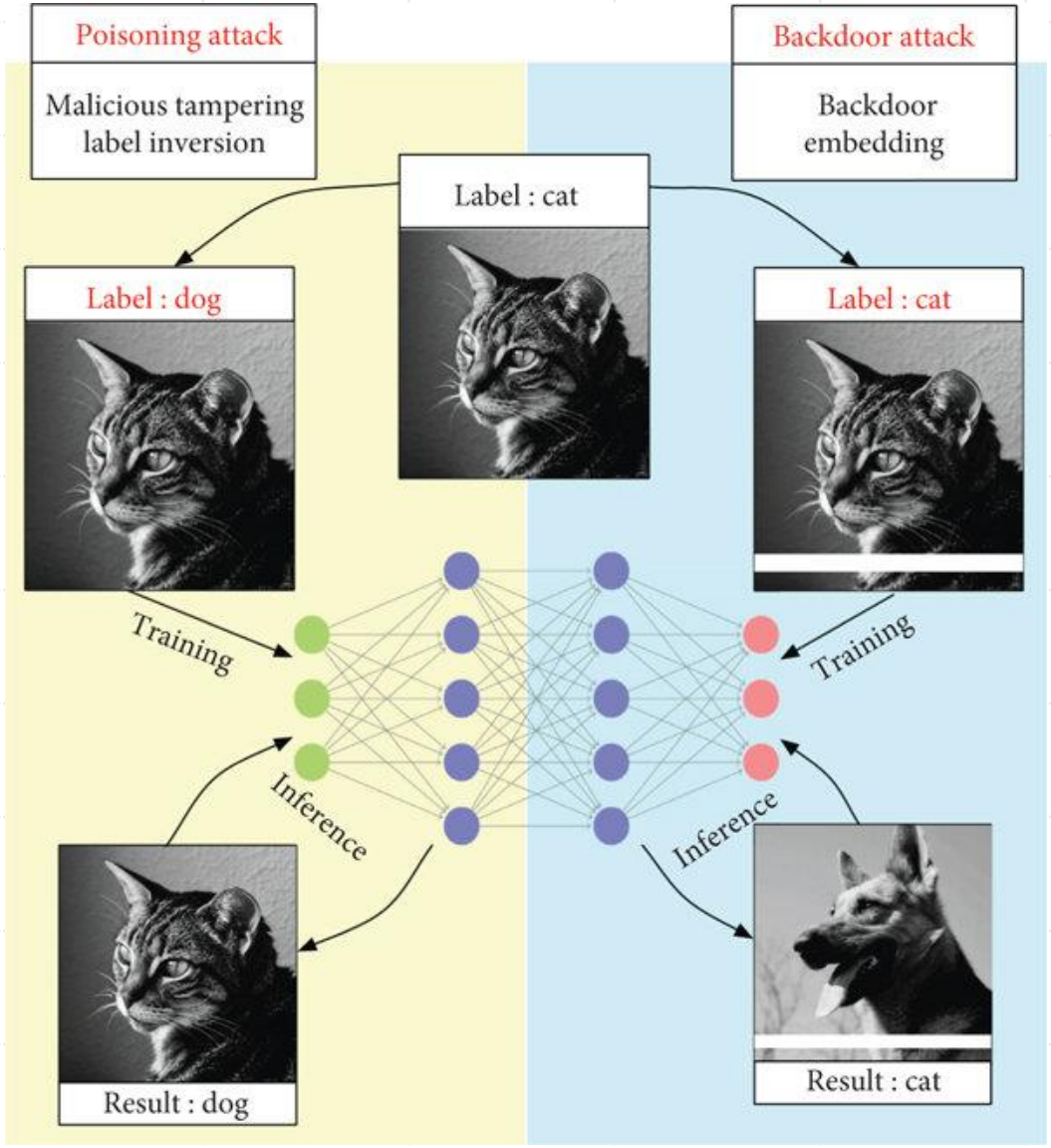
Membership Inference Attacks

Model Inversion Attacks

Model Extraction Attacks

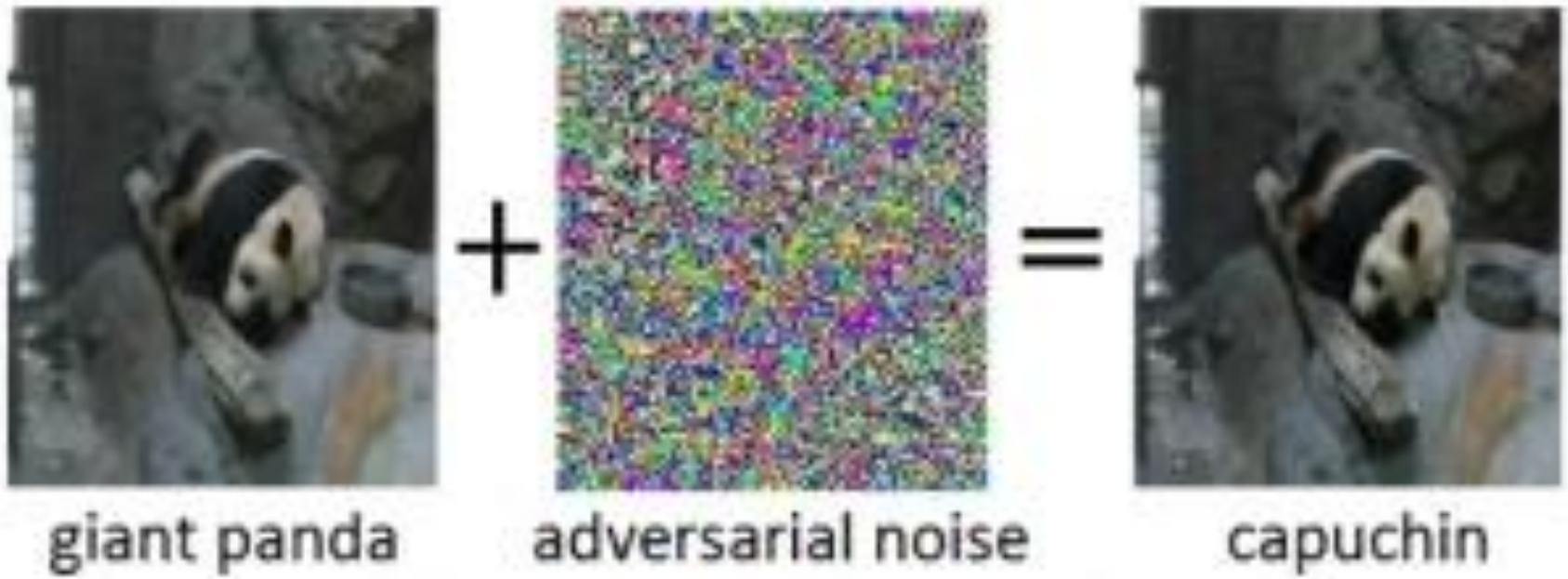
Data Poisoning

Attackers introduce malicious data into the training set to manipulate the model's behavior.



Evasion Attacks

Attackers manipulate input data in a way that alters the model's output or cause misclassification.

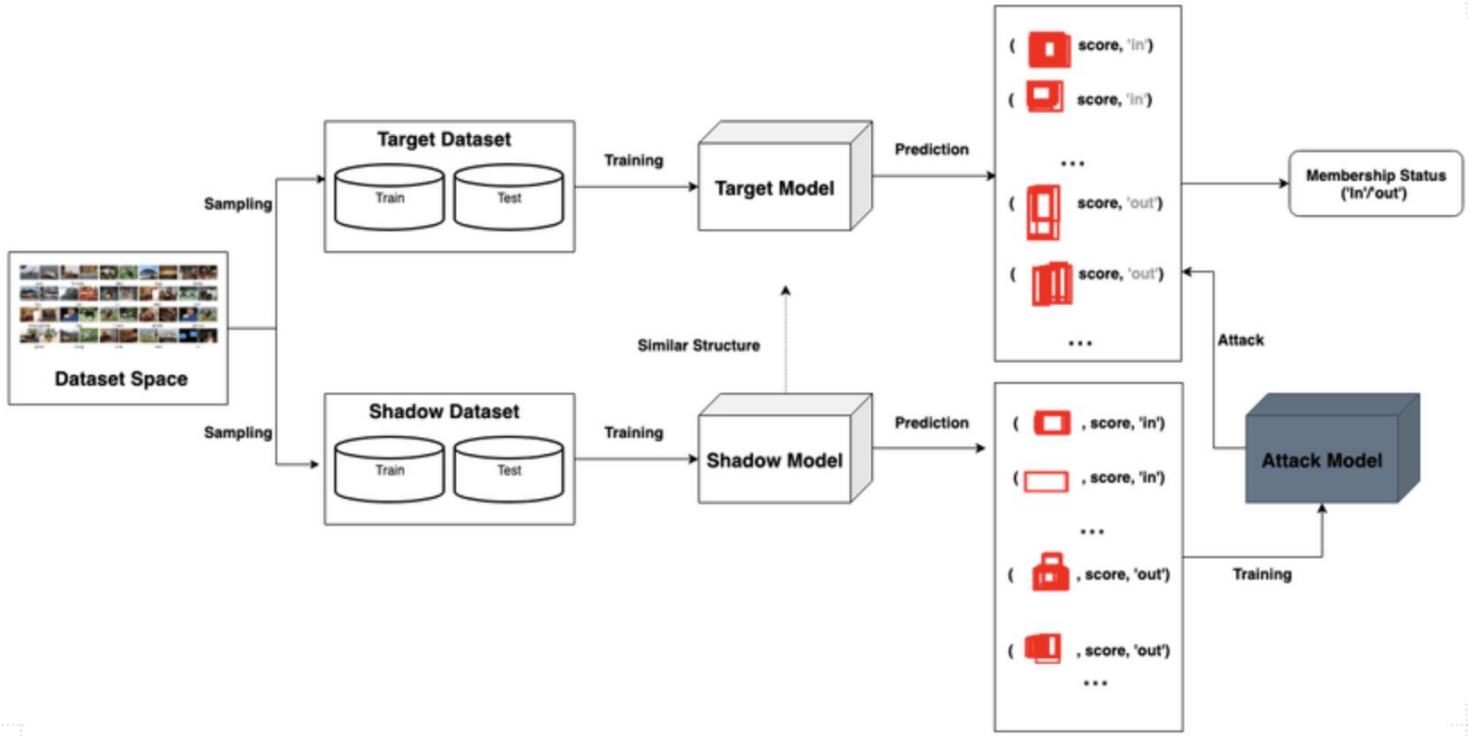


Mitigation Techniques

- **Data Validation:** techniques that can detect and remove suspicious data before training
- **Adversarial Training:** a technique improve model robustness and reduce the effect of adversarial examples
- **Model Auditing:** Regular monitoring and auditing of AI models help detect unexpected behaviors early

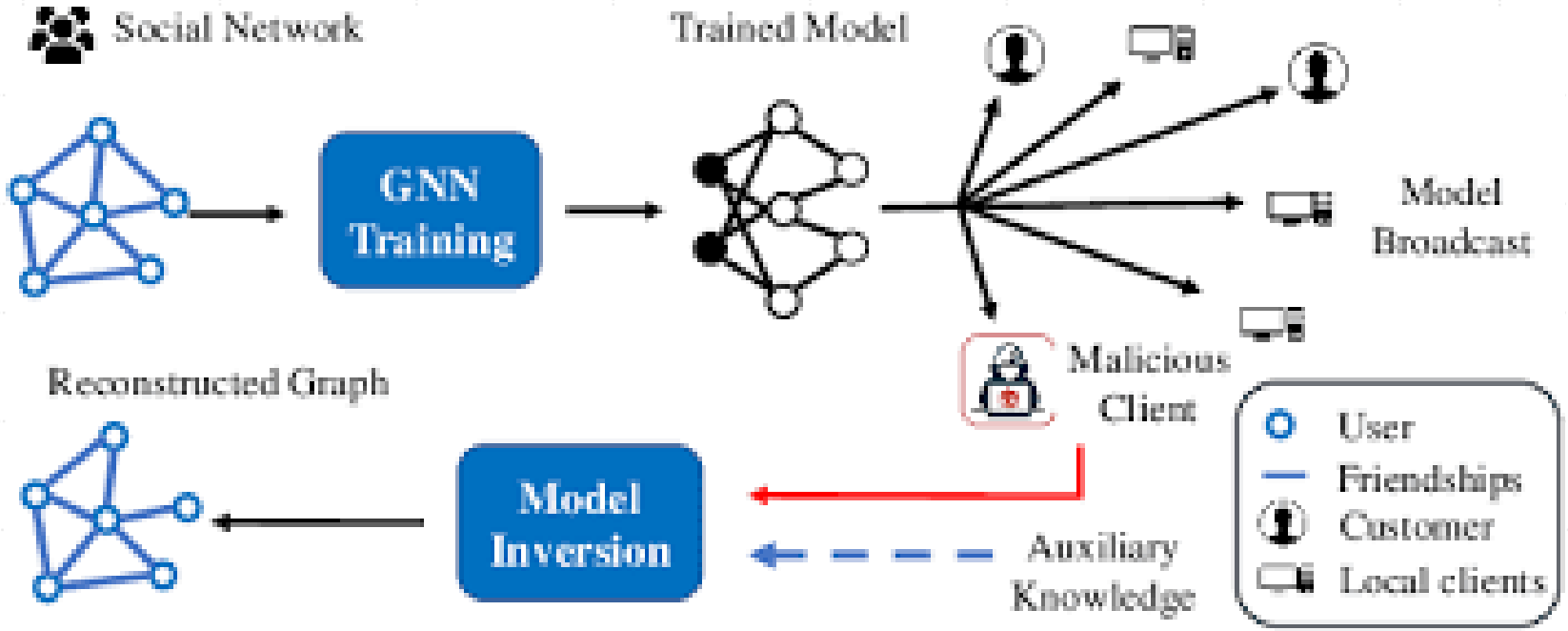
Membership Inference Attacks

Attackers infer whether a specific sample was part of the training data used by a model.



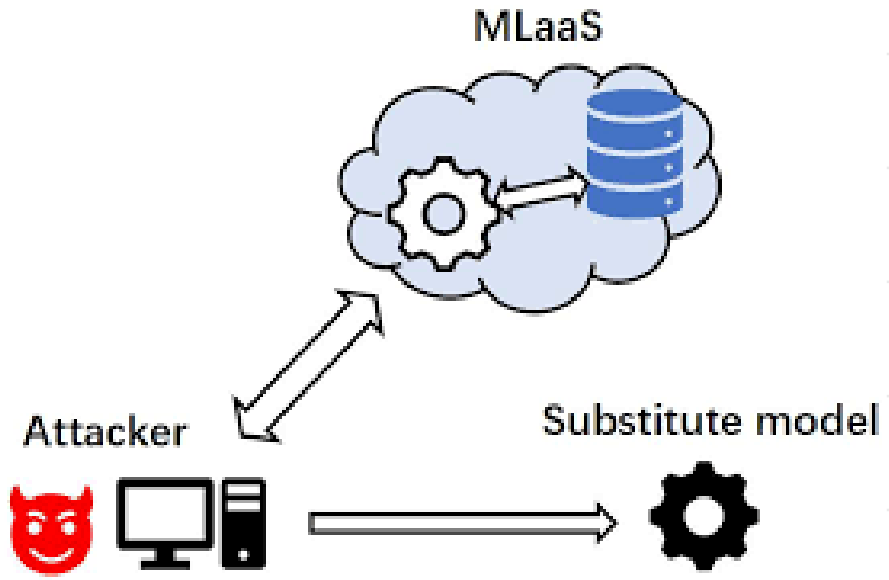
Model Inversion Attacks

Attackers attempt to reconstruct sensitive information about the training data or inputs by exploiting the model's output



Model Extraction Attacks

Attackers attempt to obtain a copy the target model by querying it and generating a substitute model.



Mitigation Techniques

- **Differential Privacy:**
- **Other Privacy-preserving techniques**

Thank You

Please send us your questions at:

vgupta@mmc.edu and

dpounds24@email.mmc.edu

Trustworthy AI: Robustness

Destiny Pounds

M.S. Student, Biomedical Data Science

Advised by Vibhuti Gupta, Ph.D.

Assistant Professor, Computer Science and Data Science

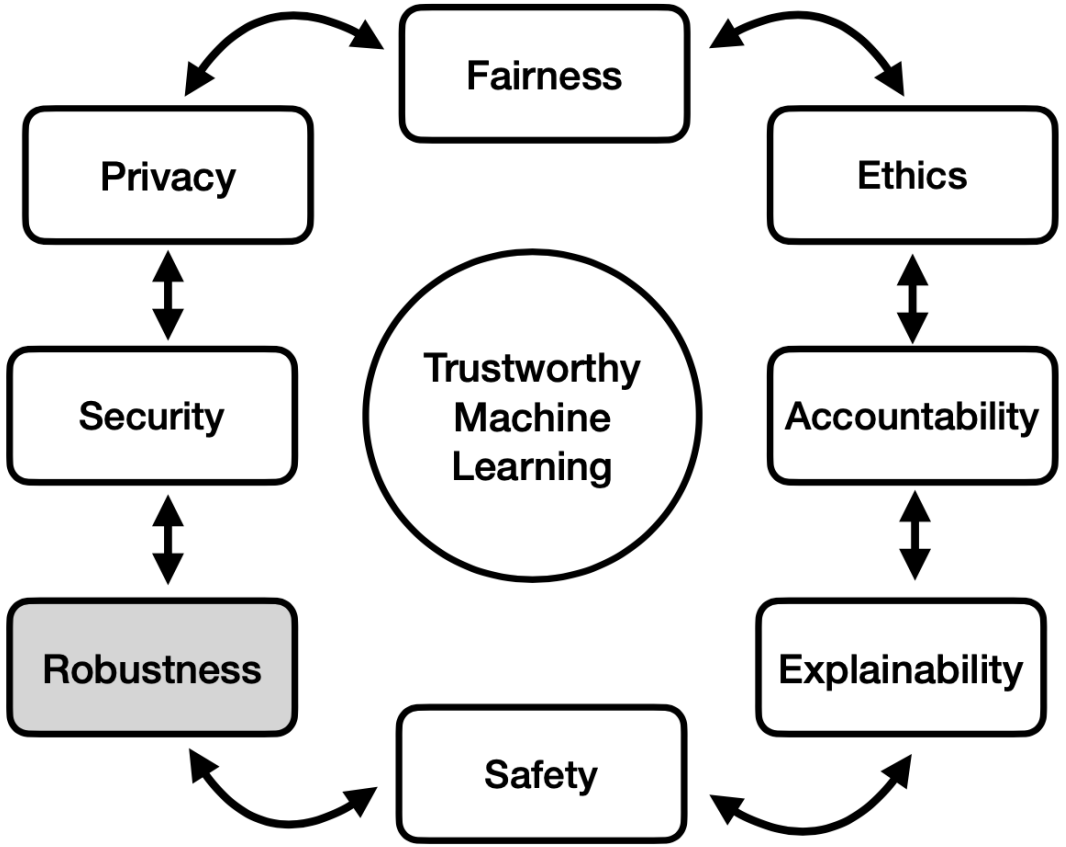
School of Applied Computational Sciences

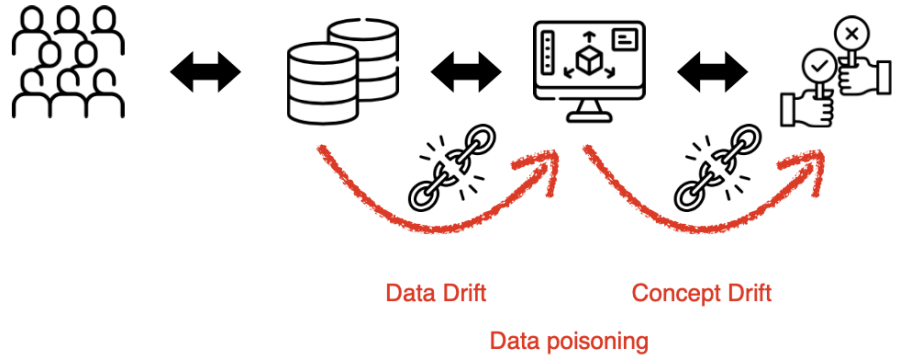
Meharry Medical College

Overview

- Robustness
- Adversarial Learning
- Potential Attacks
- Importance of Adversarial Learning

Robustness



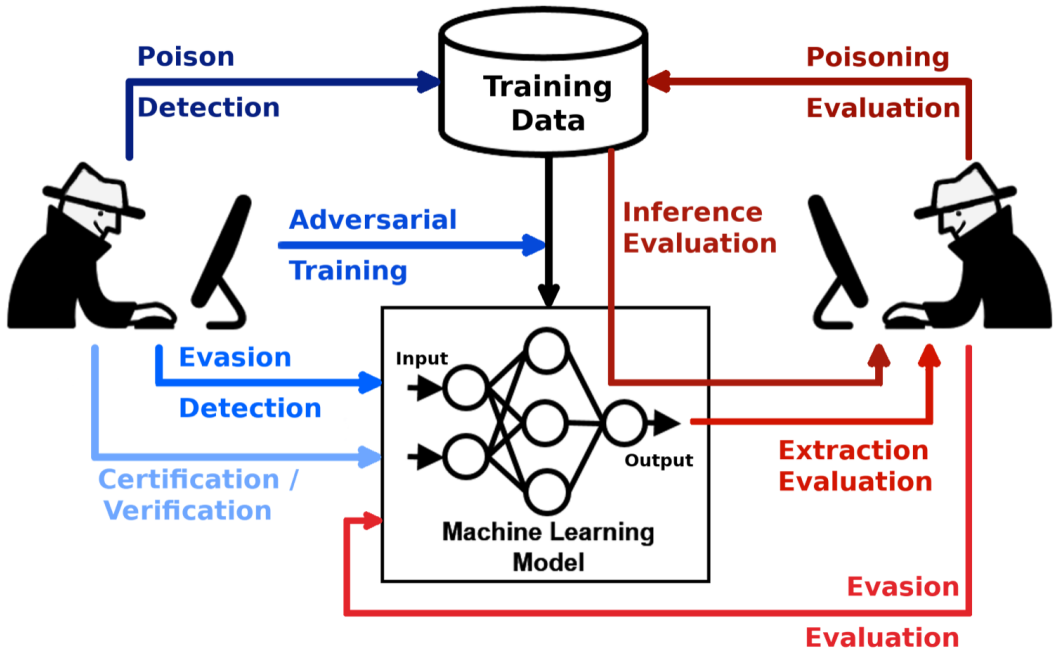


- Robustness is the property that characterizes how effective an algorithm is while being tested on a new independent dataset.

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

Adversarial Learning

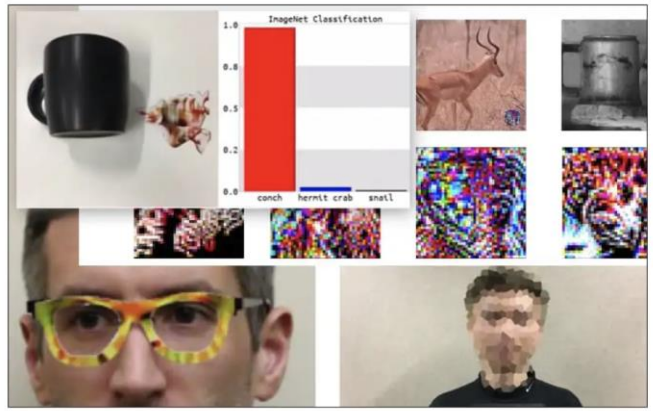
Adversarial learning involves training models to be robust against adversarial examples. These examples are intentionally designed inputs created to mislead the model into making inaccurate and wrong predictions.



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

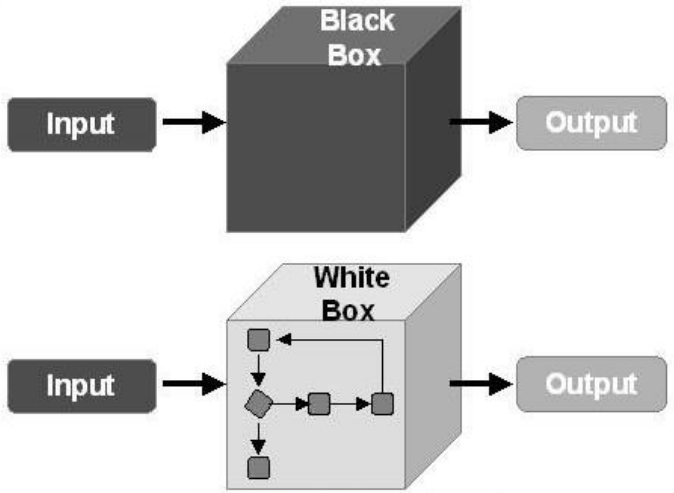
Adversarial Attacks

White Box Attacks

In a white box attack, the attacker has complete knowledge of the model and its inner workings.

Black Box Attacks

In a black box model, the attacker has limited to no knowledge of the model's internal details.

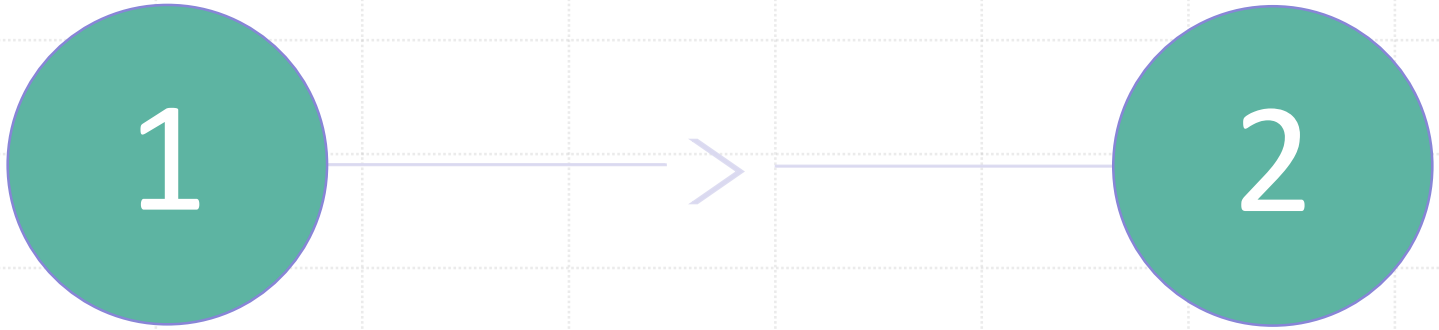


Why is adversarial learning important?

Adversarial learning improves robustness by helping the model generalize better. During training, the model is exposed to a wide range of adversarial examples which the model must adapt to.

It also helps detect weaknesses in the model and provides insights into how the model can be improved.

Incorporating adversarial learning into a machine learning model requires two steps:



Generate adversarial examples

Incorporate these examples into the training process

Thank You

Please send us your questions at:

vgupta@mmc.edu and

dpounds24@email.mmc.edu

Trustworthy AI: Safety

Destiny Pounds

M.S. Student, Biomedical Data Science

Advised by Vibhuti Gupta, Ph.D.

Assistant Professor, Computer Science and Data Science

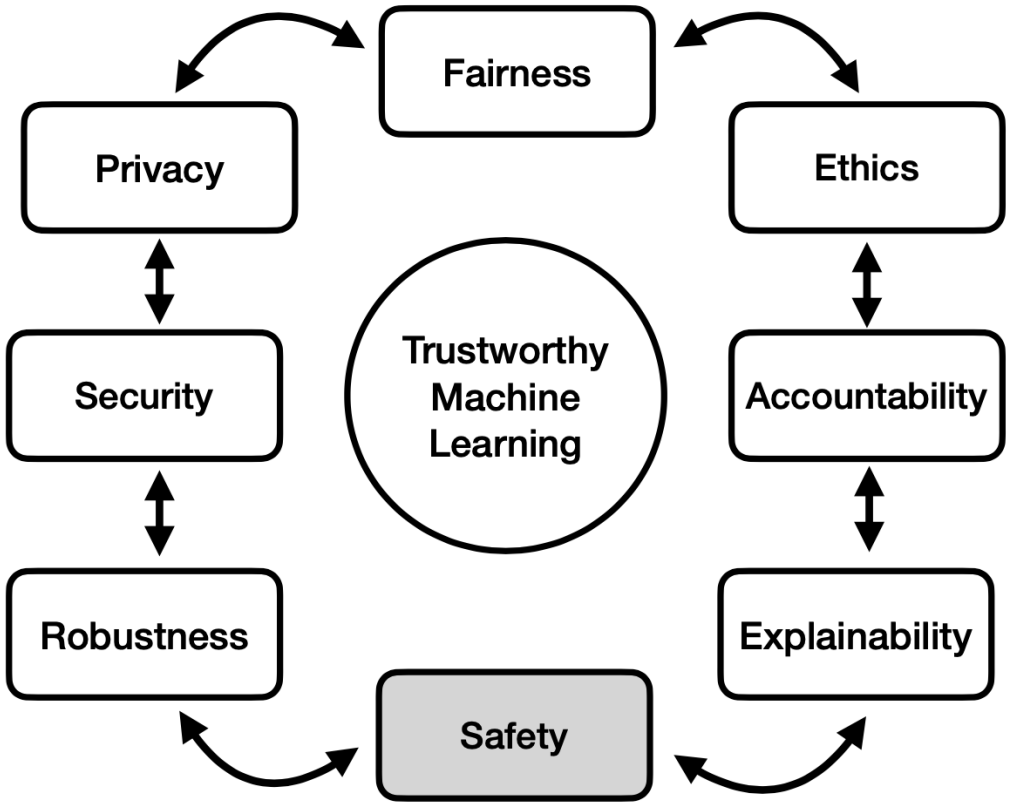
School of Applied Computational Sciences

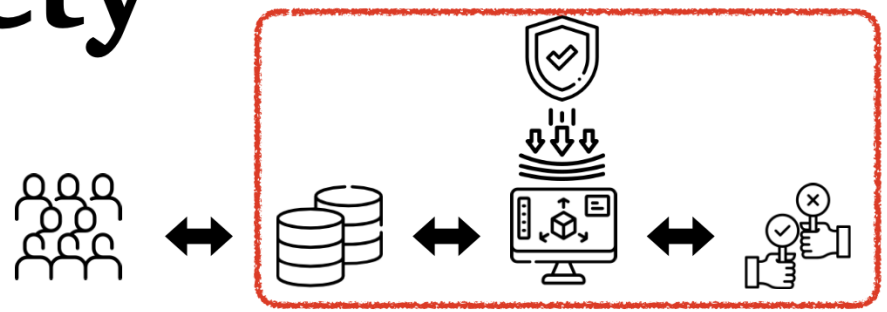
Meharry Medical College

Overview

- Defining Safety
- AI Safety
  - Robustness
  - Monitoring
    - Hazard Analysis
  - Alignment
  - Systemic Safety

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

AI Safety

- Ensuring systems can withstand hazards (Robustness)
- Identifying hazards (Monitoring)
- Reducing inherent ML system hazards (Alignment)
- Reducing systemic hazards (Systemic safety)

Robustness

- AI systems that are affected by minor disturbances could lead to safety issues.
- For example, an autonomous vehicle that fails to recognize the stop sign shown below may put passengers at risk of harm.

- The AI system should be robust that the minor disturbances do not affect accurate detection.

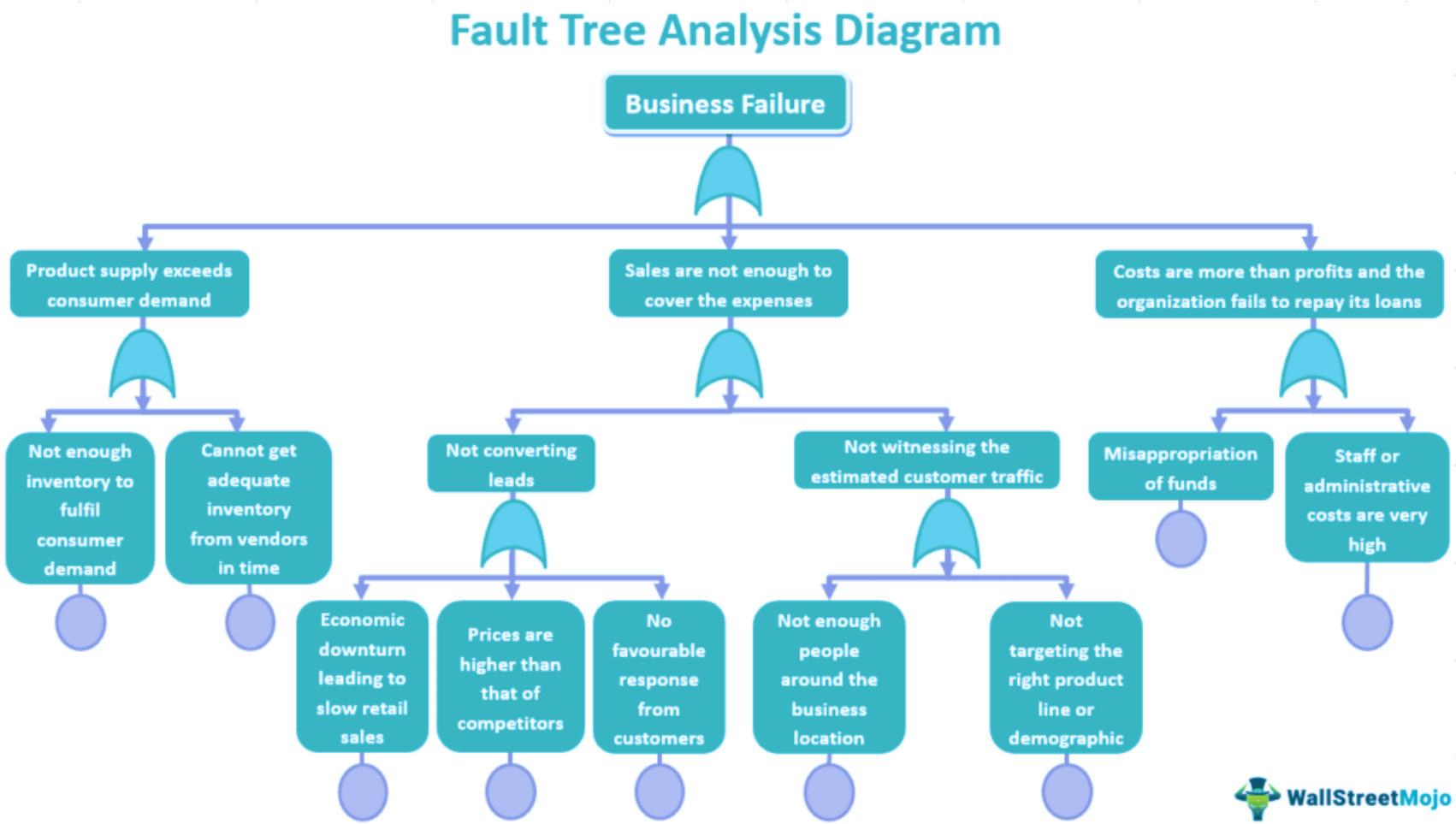
Monitoring

- Hazards are system conditions in which an accident can happen.
  - Example: an autonomous vehicle going too fast while not recognizing pedestrians
- We want to identify these hazards and the events that cause them, known as the root causes.

Hazard Analysis Techniques

- Fault Tree Analysis (FTA)
- Failure Mode and Effect Analysis (FMEA)
- Hazard and Operability Analysis (HAZOP)

Fault Tree Analysis



Failure Mode and Effect Analysis (FMEA)

Process Step or Variable or Key Input	Potential Failure Mode	Potential Effect on Customer Because of Defect	SEV	Potential Causes	OCC	Current Process Controls	DET	RPN
1. Customer Application	Application being filled out incorrectly	Application has to be resubmitted	8	Difficult to understand instructions	6	Check of application form for correct information by data entry operator	2	96
2. Data Entry	Data entered incorrectly	Customer receives checks with printing errors	4	Data entry error within a single field	6	None in place	10	240
3. Data Entry	Data entered incorrectly	Customer receives checks with printing errors	4	Information entered in wrong field	4	Self inspection	5	80

Hazard Operability Analysis (HAZOP)

This slide covers quality risk management tool such as Hazard Operability Analysis including deviation, causes, consequences, safeguards and recommendation.

Deviation	Causes	Consequences	Safeguards	Recommend
High temperature In blender	Steam heating Control malfunction	○ Feed material #1 reaches decomposition temperature ○ Violent reaction with toxic gas generation ○ Personnel exposure/injury ○ Equipment damage	○ Diverse high temp. Interlock on blender ○ Blender vented	○ Test interlock on quarterly basis ○ Add steam heating control to monthly PM

Alignment

Alignment refers to the concept of designing the right objective function for a task, so that the task is encoded in alignment with how humans intend the AI to perform the task. This reduces hazards by specifying the right requirements for the system.

Example:

A sidewalk robot with the objective function to reach a goal fast might drive at dangerously high speeds and might endanger cyclists in bike lanes, unless speed limits and not interfering with other traffic is embedded in the objective.

Task met, but not as the user intended.

Alignment Challenges

- Alignment is not an easy task as defining the objective function that takes all requirements into count require understanding and expressing requirements precisely, anticipating potential loopholes.

- Through system learning, the AI usually learns loopholes.
  - Examples:
    - AI pausing a video game indefinitely to avoid losing
    - A racing robot changing its path to cross the finish line quicker

Systemic Safety

- AI systems interact with their environment and other systems to explore strategies that ensure their safety and reliability.

Thank You

Please send us your questions at:

vgupta@mmc.edu and

dpounds24@email.mmc.edu

Trustworthy AI: Explainability

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

- Defining Explainability
- Model Understanding
  - Examples
  - Benefits
- Approaches to Model Understanding
  - Interpretable Models
  - Post-hoc Explainability
    - Local Explanations
    - Global Explanations

Explainable Artificial Intelligence (XAI)

- Explainability of an AI model describes the extent to which human-users can comprehend and trust the results and output created by the model.

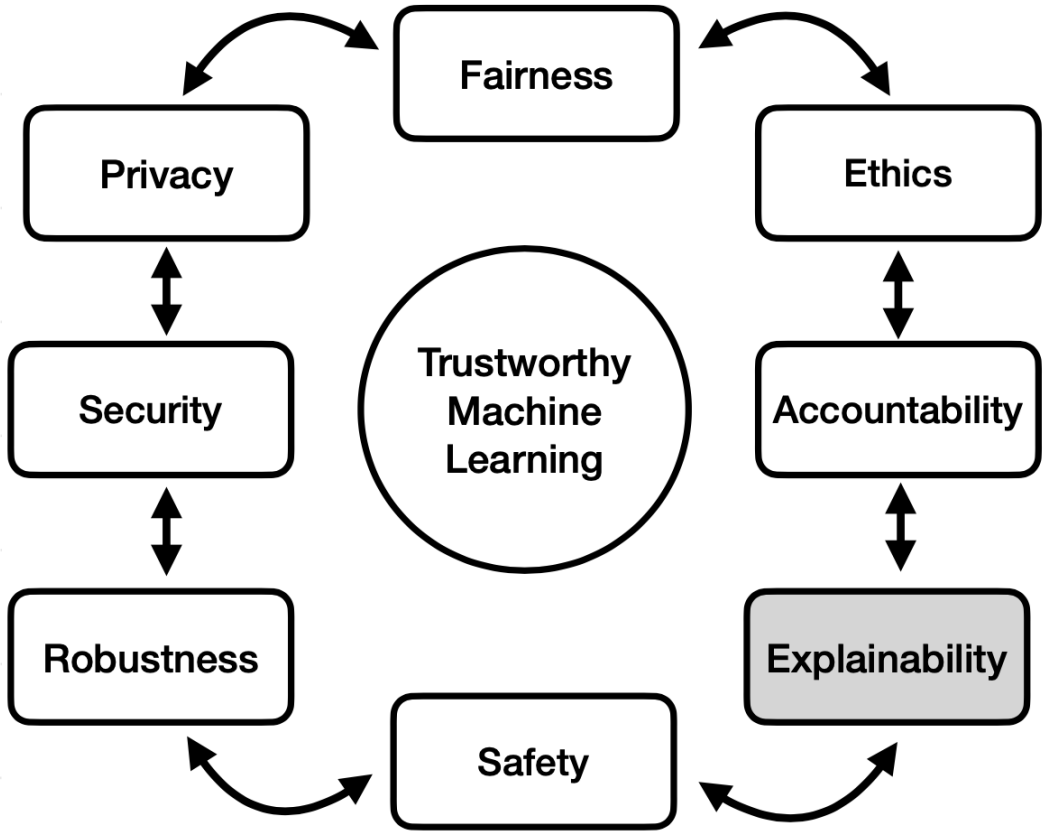


Image from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Overview of Predictive Modeling Process

Explainable AI requires model understanding.

Example: Why Model Understanding?

Input

Predictive Model

Prediction = Siberian Husky

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

Example: Why Model Understanding?

Input

Predictive Model

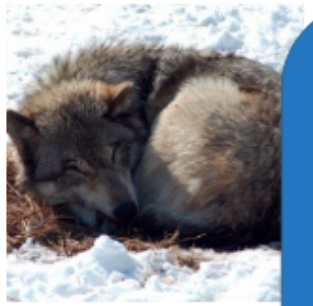
Model Understanding

Prediction = Siberian Husky

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

Example: Why Model Understanding?

Input



This model is
incorrect
make
on!! Let
model

Model understanding facilitates debugging.

Predictive Model

Prediction = Siberian Husky

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

Example: Why Model Understanding?

Defendant Details

Predictive Model

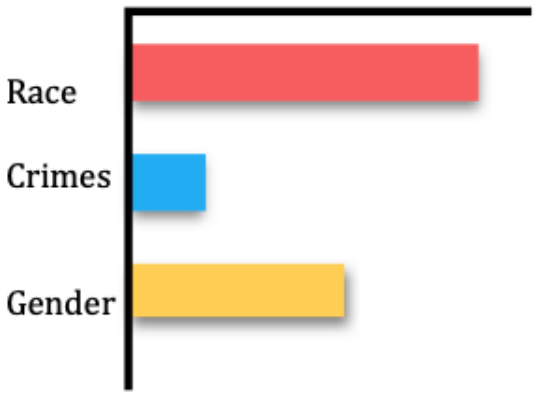
Prediction = Risky to Release

Example: Why Model Understanding?

Defendant Details

Predictive Model

Model Understanding



Prediction = Risky to Release

This prediction is biased. Race and gender are being used to make the prediction!!

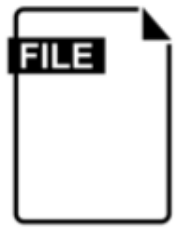
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Example: Why Model Understanding?

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

Example: Why Model Understanding?

Loan Applicant Details



Predictive Model

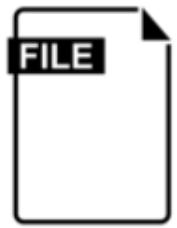
Prediction = Denied Loan

Loan Applicant

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

Example: Why Model Understanding?

Loan Applicant Details



Predictive Model

Model Understanding

Increase salary by 50K + pay credit card bills on time for next 3 months to get a loan

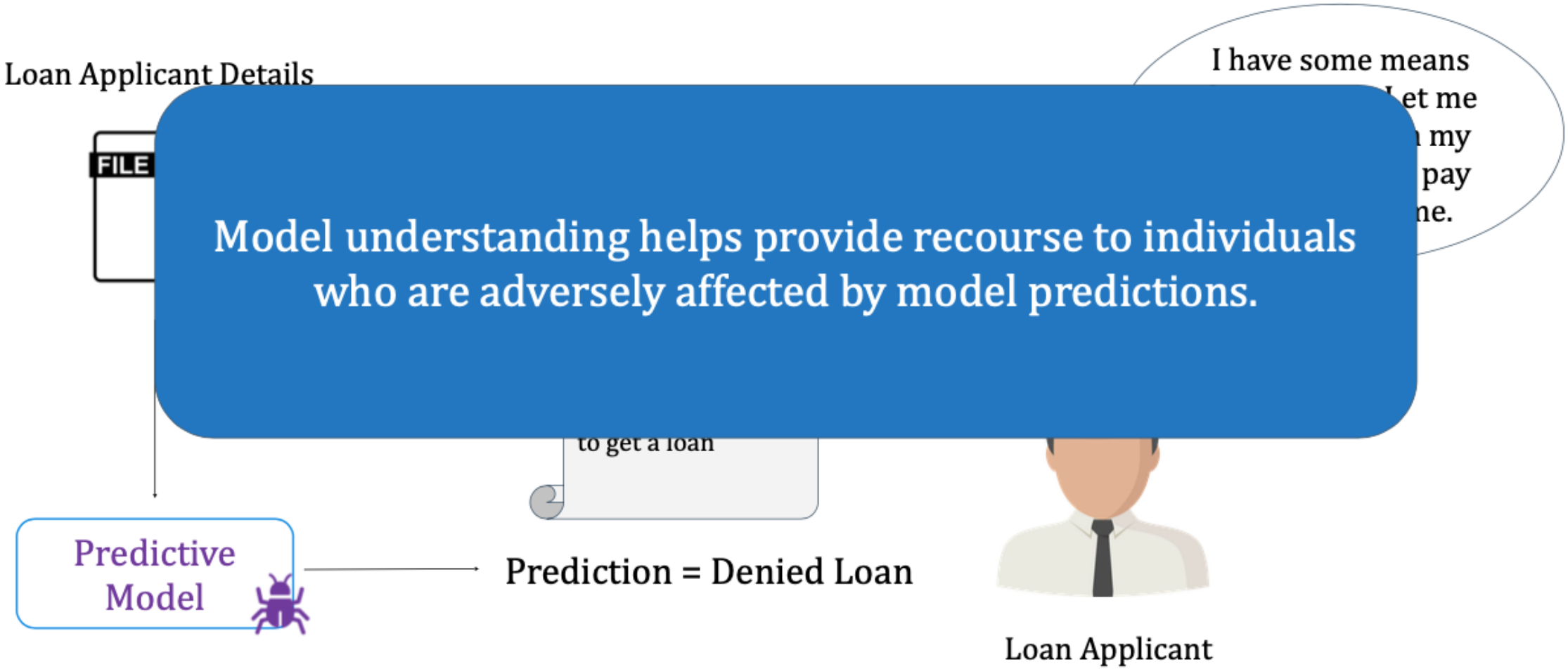
Prediction = Denied Loan

I have some means for recourse. Let me go and work on my promotion and pay my bills on time.

Loan Applicant

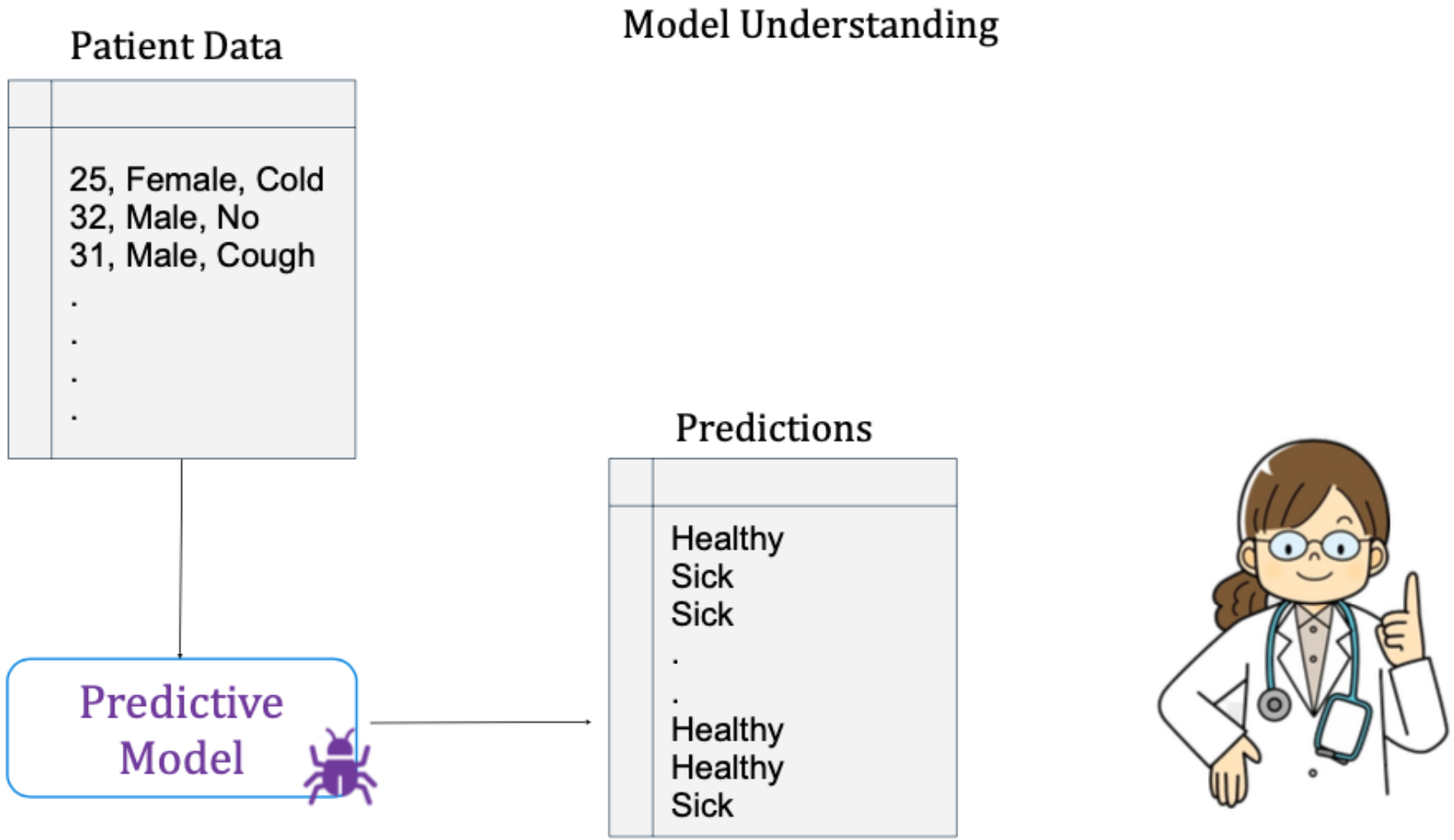
Adapted from Farnadi, G. (2022). *Trustworthy Machine Learning* [Slides]. HEC Montréal

Example: Why Model Understanding?



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

Example: Why Model Understanding?



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

Example: Why Model Understanding?

Patient Data

25, Female, Cold
32, Male, No
31, Male, Cough
.
.
.
.

Model Understanding

If gender = female,
  if ID_num > 200, then sick

If gender = male,
  if cold = true and cough = true, then sick

Predictive Model

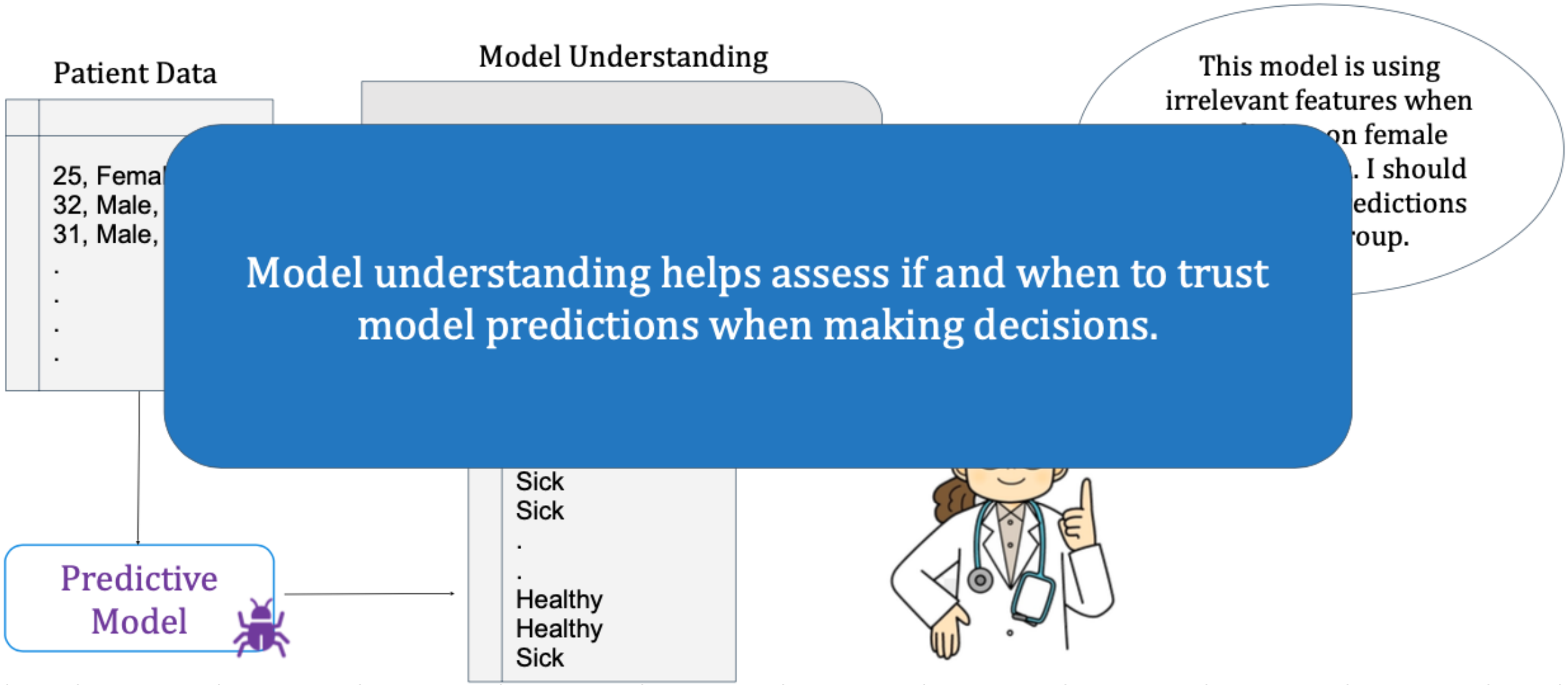
Predictions

Healthy
Sick
Sick
.
.
Healthy
Healthy
Sick

This model is using irrelevant features when predicting on female subpopulation. I should not trust its predictions for that group.

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

Example: Why Model Understanding?



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

Example: Why Model Understanding?

Patient Data

25, Female, Cold
32, Male, No
31, Male, Cough
.
.
.
.

Model Understanding

If gender = female,
  if ID_num > 200, then sick

If gender = male,
  if cold = true and cough = true, then sick

Predictive Model

Predictions

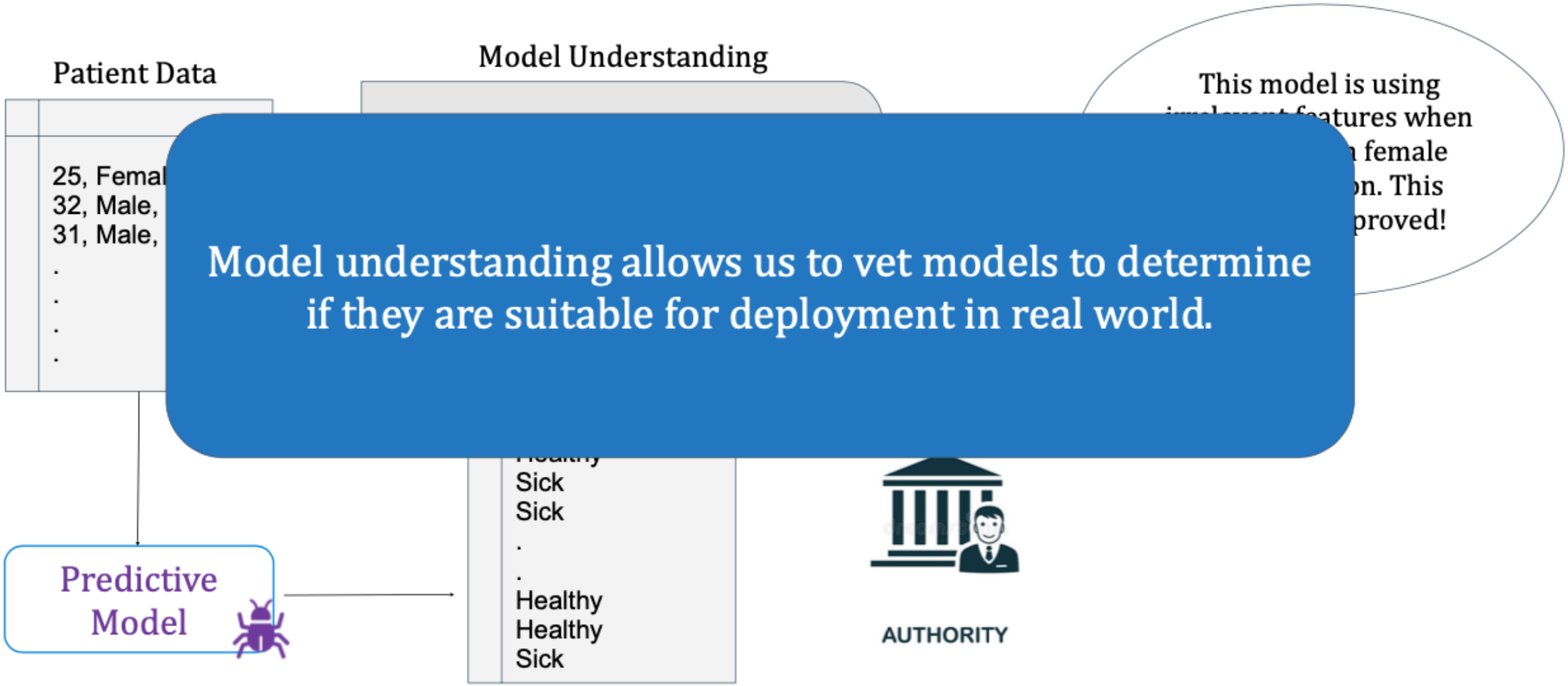
Healthy
Sick
Sick
.
.
Healthy
Healthy
Sick

AUTHORITY

This model is using irrelevant features when predicting on female subpopulation. This cannot be approved!

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

Example: Why Model Understanding?



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

Summary: Benefits of Model Understanding

Debugging

Bias Detection

Recourse

Assess Trustworthiness of Predictions

Vet Models for Deployment

Achieving Model Understanding

Approach #1:

Build
inherently interpretable
(white box) models.

Approach #2:

Explain pre-built (black-box) models in a post-hoc manner.

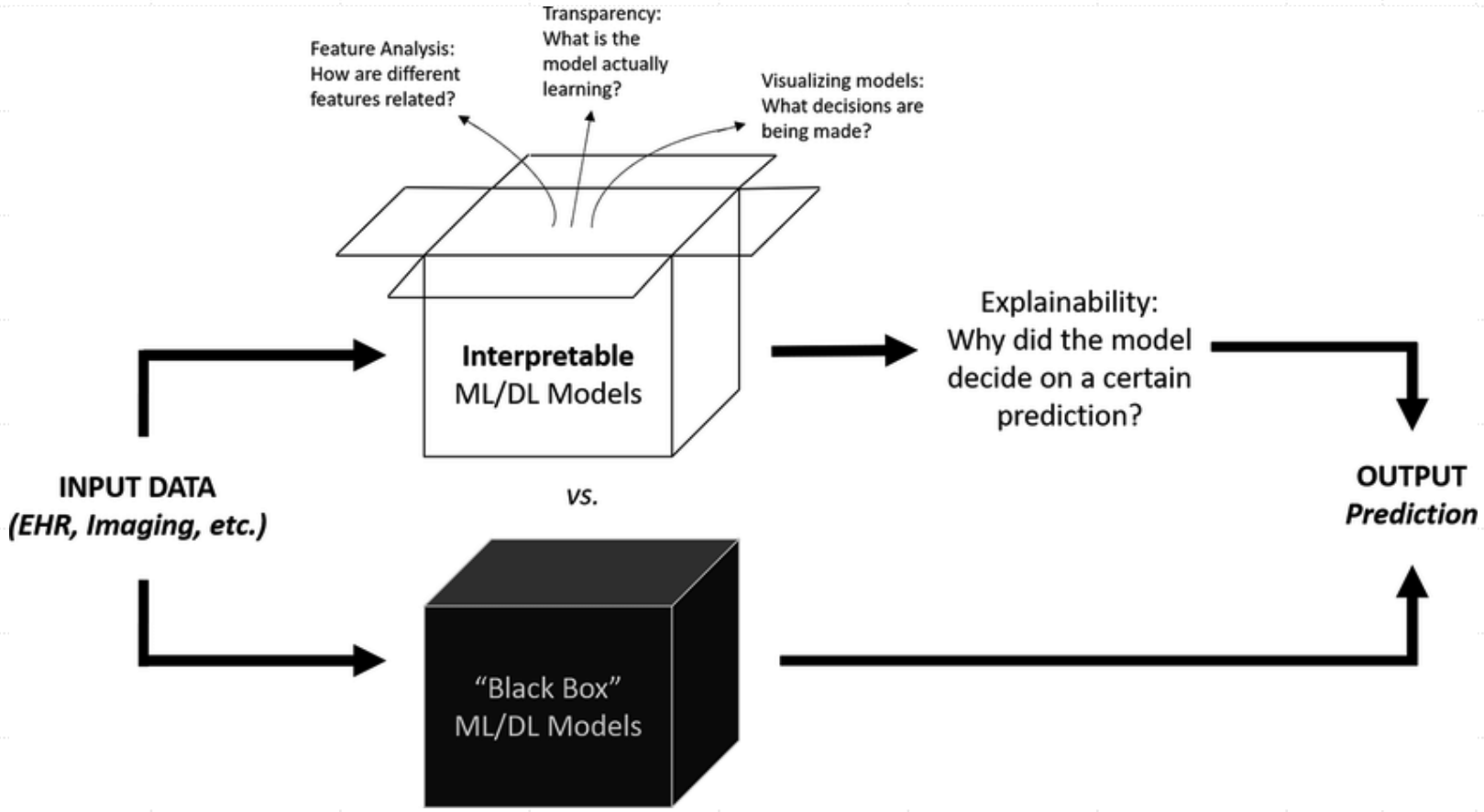
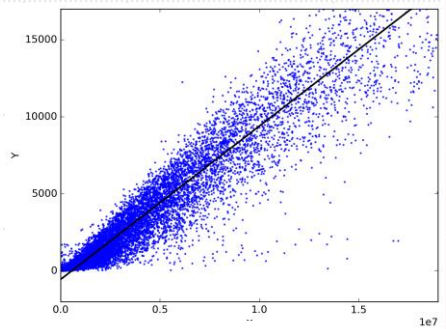
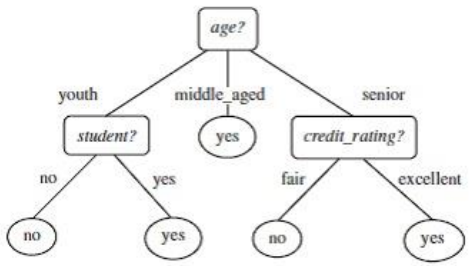


Image from Hui, Aaron & Ahn, Shawn & Lye, Carolyn & Deng, Jun. (2022). Ethical Challenges of Artificial Intelligence in Health Care: A Narrative Review. Ethics in Biology, Engineering and Medicine: An International Journal. 12. 10.1615/EthicsBiologyEngMed.2022041580.

Inherently Interpretable Models

- These models are interpretable by design. Understanding of the model is clear before receiving results.



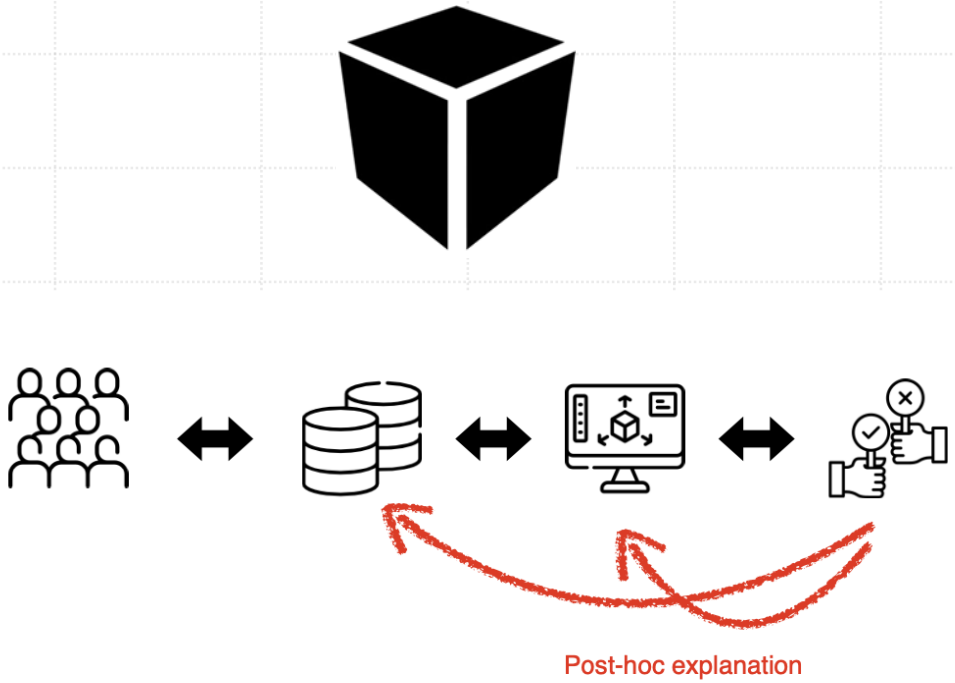


The HEART Score for Chest Pain Patients in the ED		
History	• Highly Suspicious • Moderately Suspicious • Slightly or Non-Suspicious	• 2 points • 1 point • 0 points
ECG	• Significant ST-Depression • Nonspecific Repolarization • Normal	• 2 points • 1 point • 0 points
Age	• ≥ 65 years • > 45 - < 65 years • ≤ 45 years	• 2 points • 1 point • 0 points
Risk Factors	• ≥ 3 Risk Factors or History of CAD • 1 or 2 Risk Factors • No Risk Factors	• 2 points • 1 point • 0 points
Troponin	• ≥ 3 x Normal Limit • > 1 - < 3 x Normal Limit • ≤ Normal Limit	• 2 points • 1 point • 0 points
Risk Factors: DM, current or recent (<one month) smoker, HTN, HLP, family history of CAD, & obesity		
Score 0 – 3: 2.5% MACE over next 6 weeks → Discharge Home **Score 4 – 6:** 20.3% MACE over next 6 weeks → Admit for Clinical Observation **Score 7 – 10:** 72.7% MACE over next 6 weeks → Early Invasive Strategies		

Examples

- Linear Models
- Shallow Decision Tree
- Rule Based Models
- Risk Scores

Post-hoc Explainability

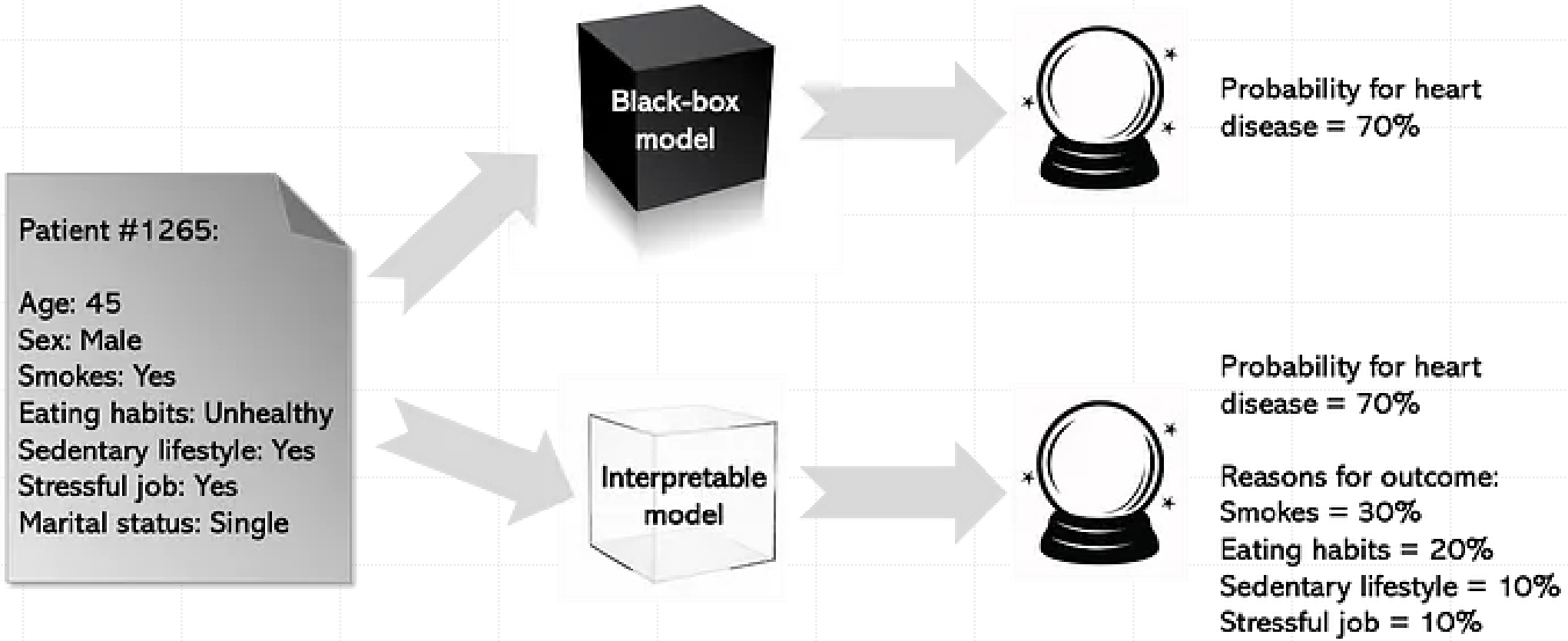


Building an inherently interpretable model is not always possible and we are left with a black box. In that case, models can be interpreted post-hoc.

Post-hoc explanation occurs after execution of the model.

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

White Box vs. Black Box Example



Approaches for Post-hoc Explainability

Local Explanations

Local explanations focus on the data and provide explanations for individual outcomes. Therefore, they provide trust for model outcomes.

Global Explanations

Global explanations focus on the model and provide an understanding of the decision process. Therefore, they provide a sense of understanding to how the model works.

Thank You

Please send us your questions at:

vgupta@mmc.edu and

dpounds24@email.mmc.edu

Trustworthy AI: Accountability

Destiny Pounds

M.S. Student, Biomedical Data Science

Advised by Vibhuti Gupta, Ph.D.

Assistant Professor, Computer Science and Data Science

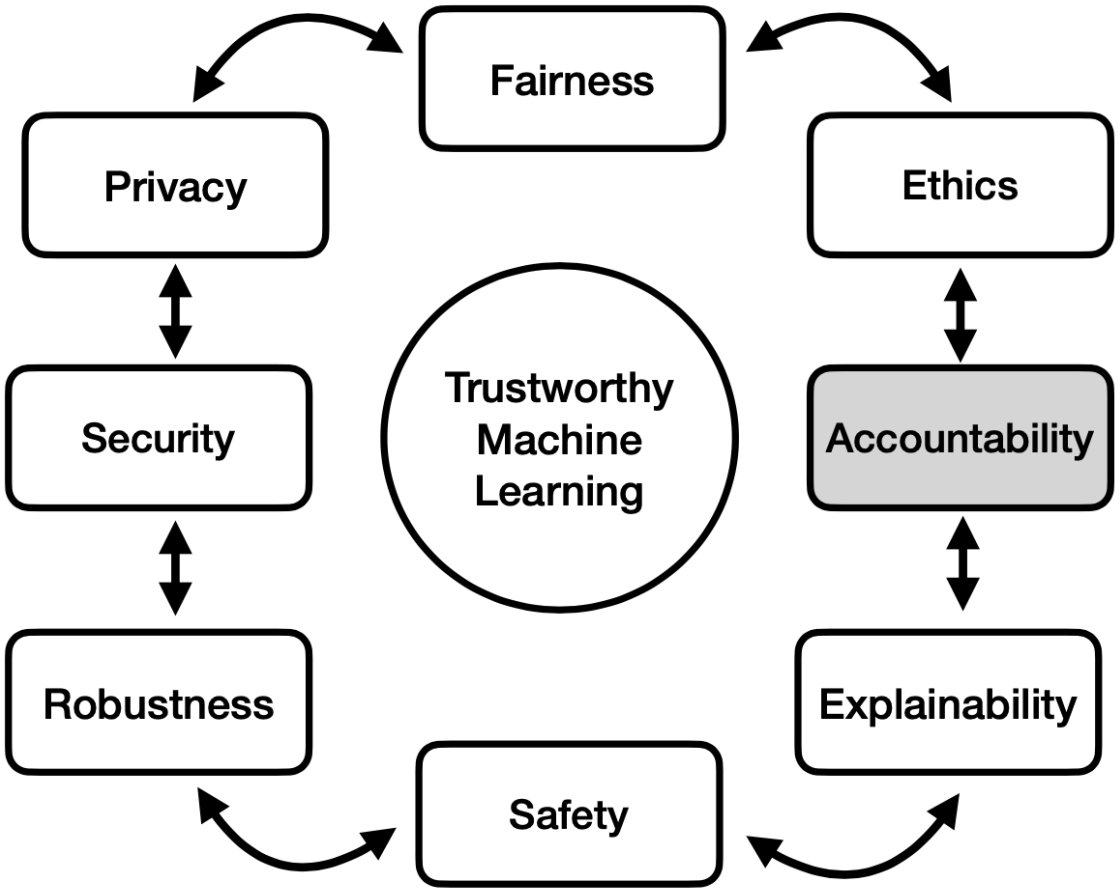
School of Applied Computational Sciences

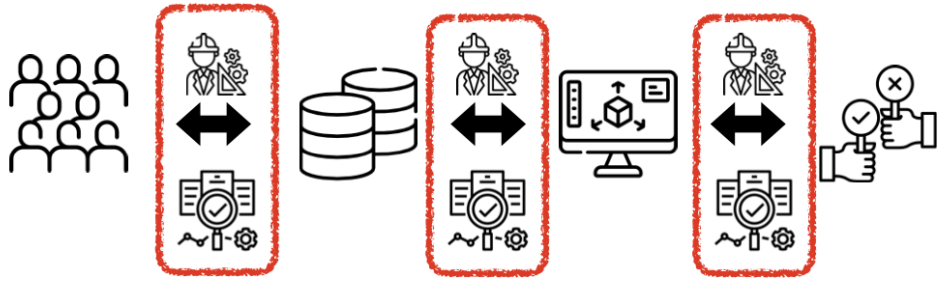
Meharry Medical College

Overview

- Defining Accountability
- Roles
  - System Designers
  - Decision Makers
  - System Developers
  - System Auditors
  - End Users
- Audits
  - External
  - Internal

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

Roles

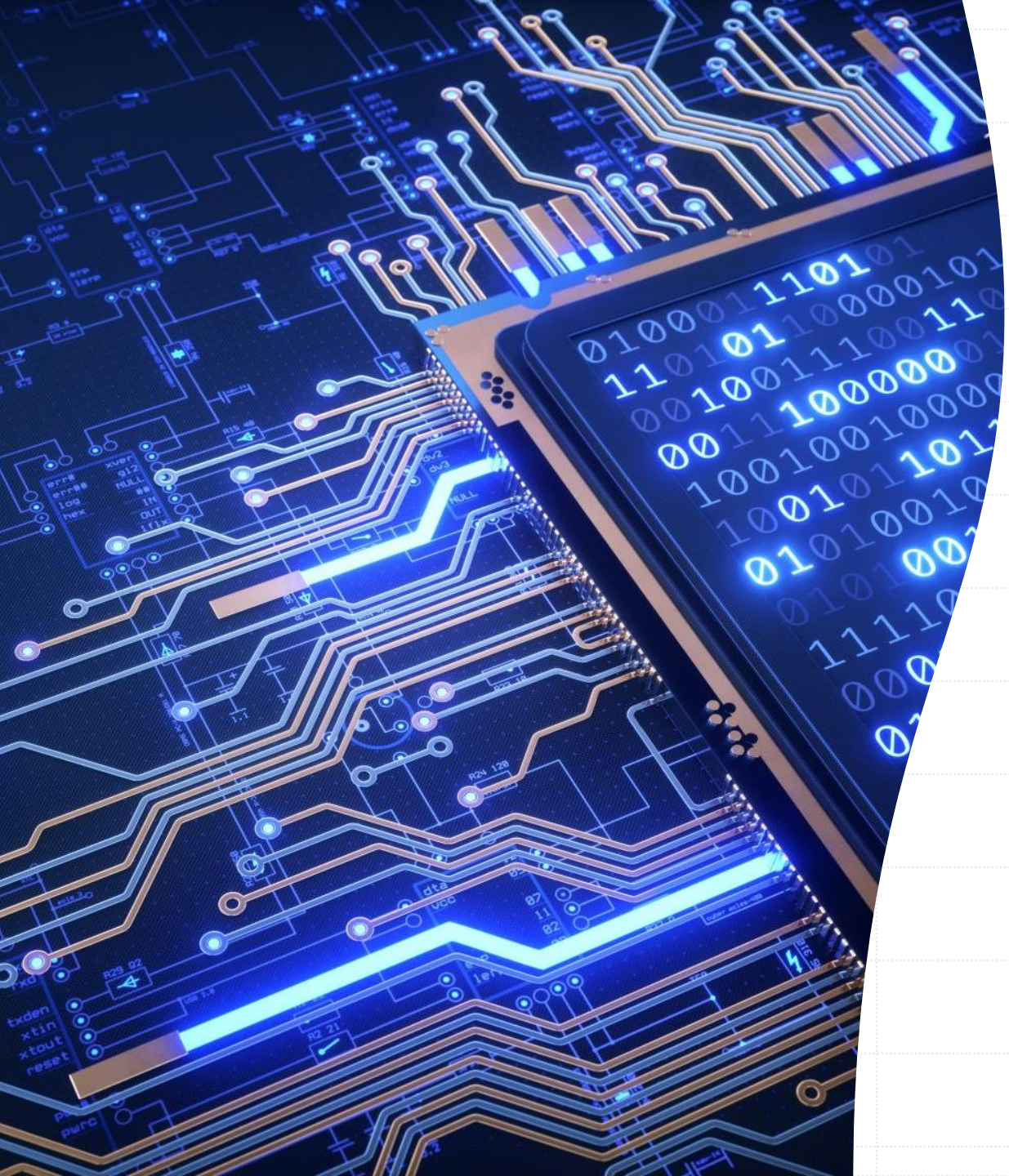
SYSTEM DESIGNERS

DECISION MAKERS

SYSTEM DEPLOYERS

SYSTEM AUDITORS

END USERS

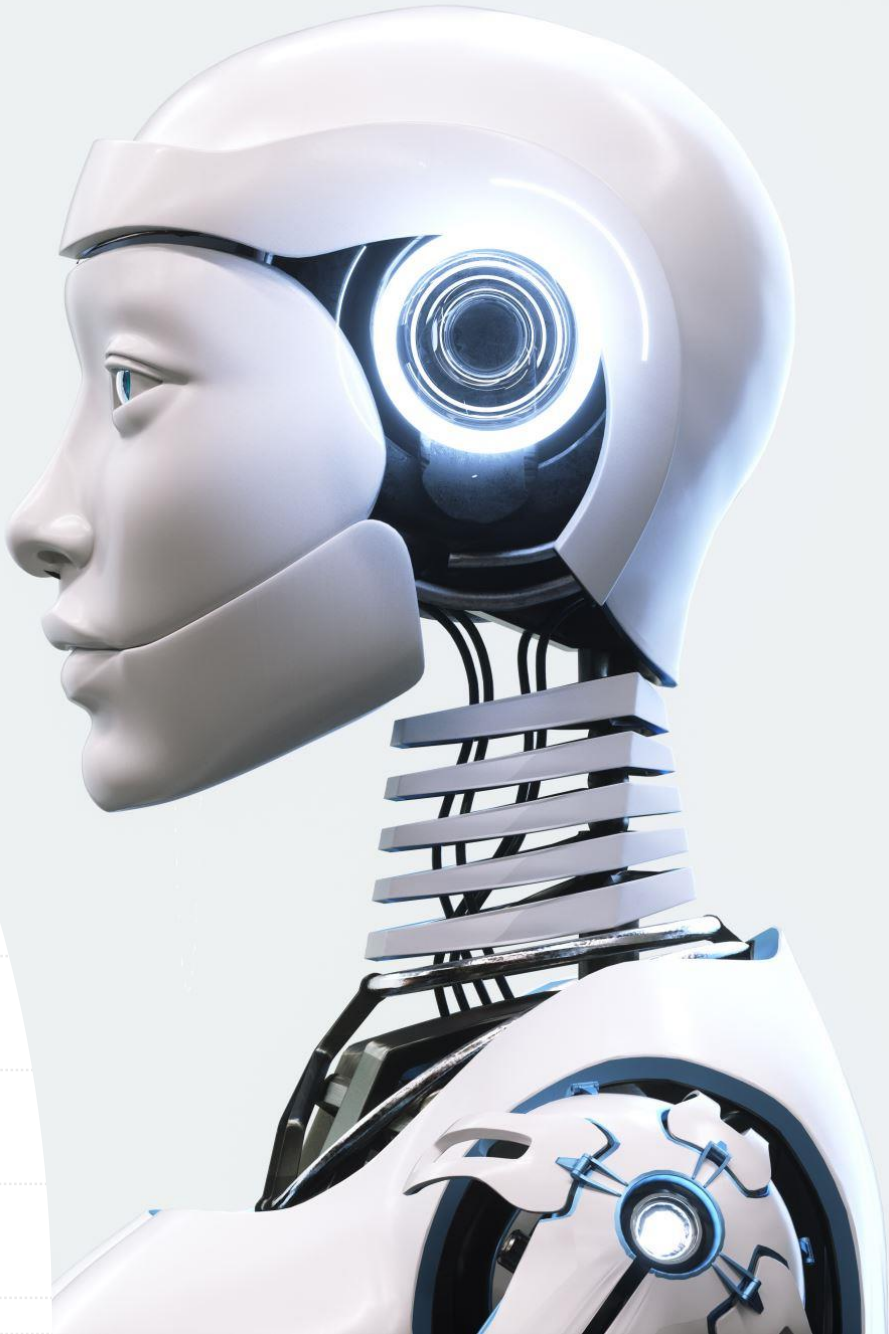


System designers

- System designers are the designers of the AI system. Their job is to design the AI system to user requirements so that it is transparent and explainable.
  System designers also
  provide deployment instructions and user guidelines.

Decision Makers

- Decision makers have the right to build an AI system and determine what AI system should be adopted. They should be fully aware of the benefits and risks of the candidate AI systems and take all requirements and regulations into consideration.

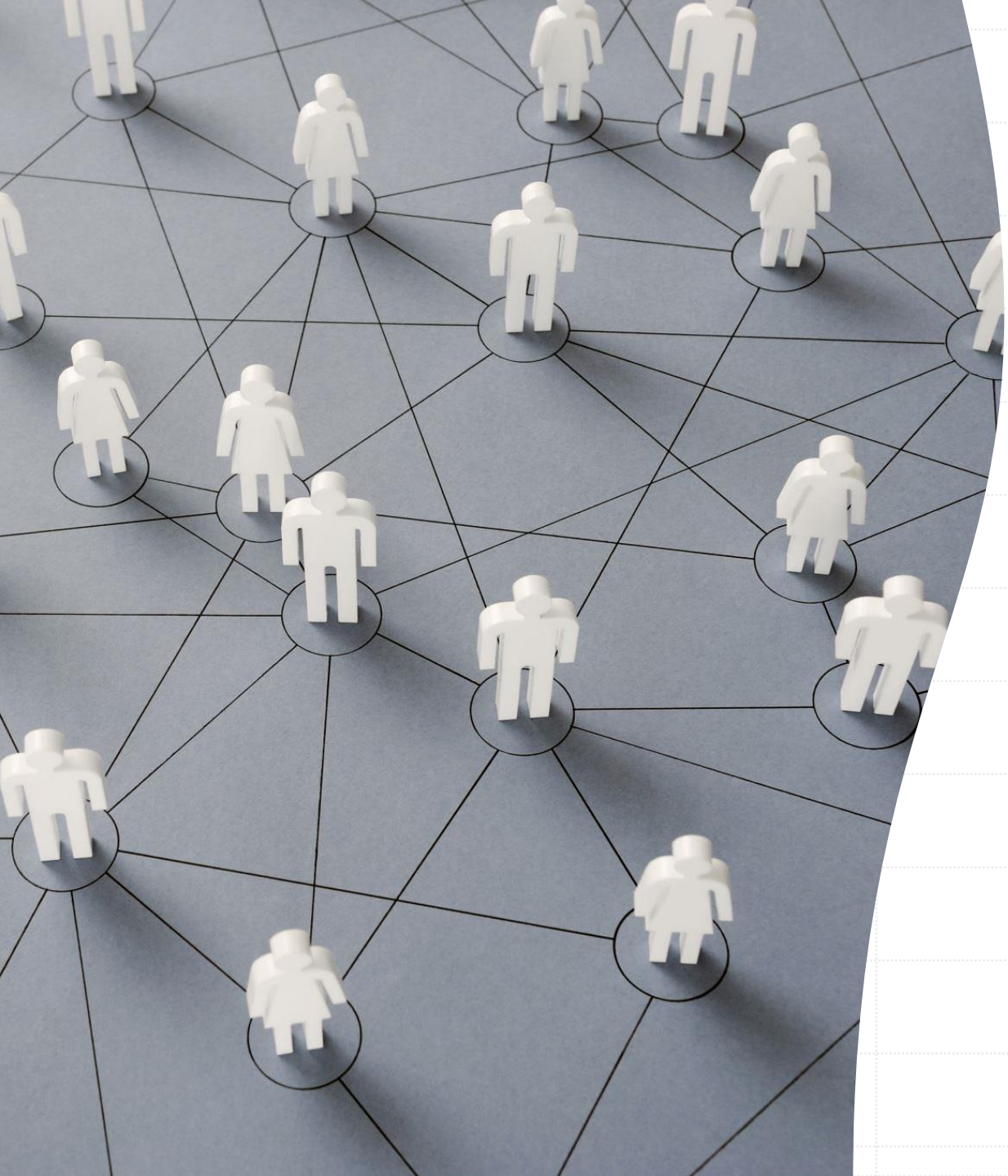


System Deployers

- System deployers oversee the deployment of the AI system. Their job is to follow deployment instructions and ensure the system is deployed appropriately.

System Auditors

- System auditors are responsible for system auditing. Auditors are expected to provide comprehensive assessments of the AI system.



End Users

- End users are the practical operators of the AI system. End users are expected to follow the user guidelines and report any issues to deployers and designers in a timely manner.

Audits

- An audit is an evaluation of conformance of AI systems and processes to applicable regulations, standards, guidelines, plans, specifications and procedures.

- Audit Types
  - Internal audit
  - External audit

Internal Audit

- An internal audit is conducted by people inside same organization as the system designer or system deployer.
- The auditor will have access to large amounts of internal data.
- This audit can be done before deployment; therefore, the decision maker can utilize audit results.
- However, the auditor shares same interests as the audited which makes objective report more challenging.

External Audit

- An external audit is an audit conducted by a third party with no common interest.
- This type can easily offer a comprehensive and objective report due to lack of potential bias.
- The auditor cannot access all the important internal data in AI system.
- The audit must be done after deployment, so it be costly to adjust as needed.

Thank You

Please send us your questions at:

vgupta@mmc.edu and

dpounds24@email.mmc.edu