BEYOND THE HYPE

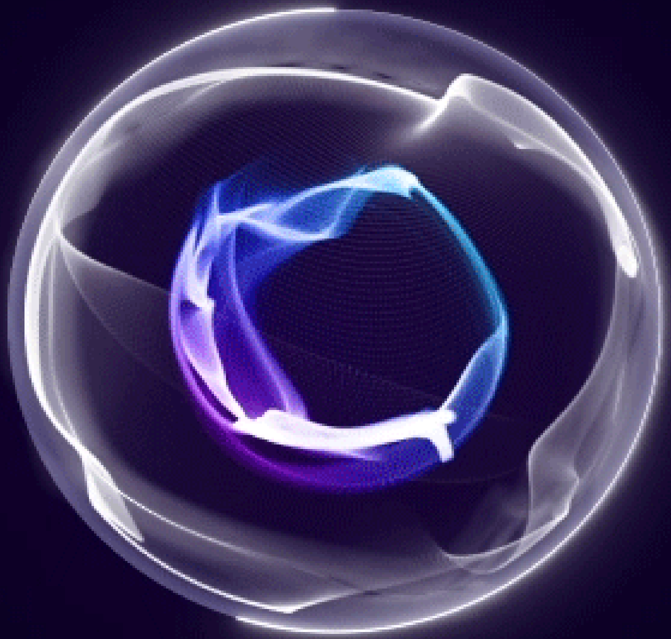
Unraveling the Myths, Realities, & Governance of Artificial Intelligence

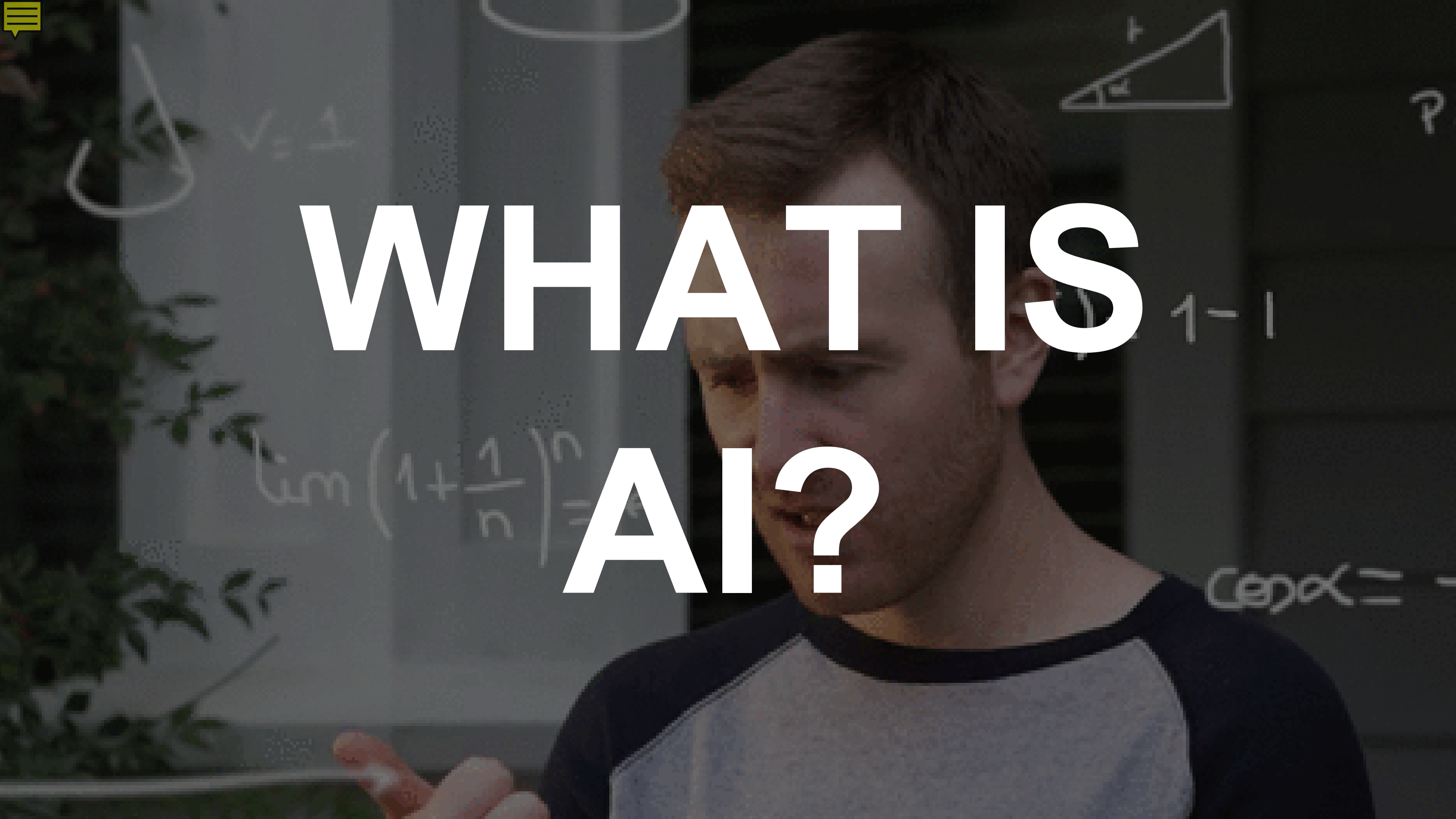
Brandie Nonnecke, PhD
Director, CITRIS Policy Lab
Assoc. Research Professor, Goldman School of Public Policy
UC Berkeley
@BNonnecke | nonnecke@berkeley.edu

2024

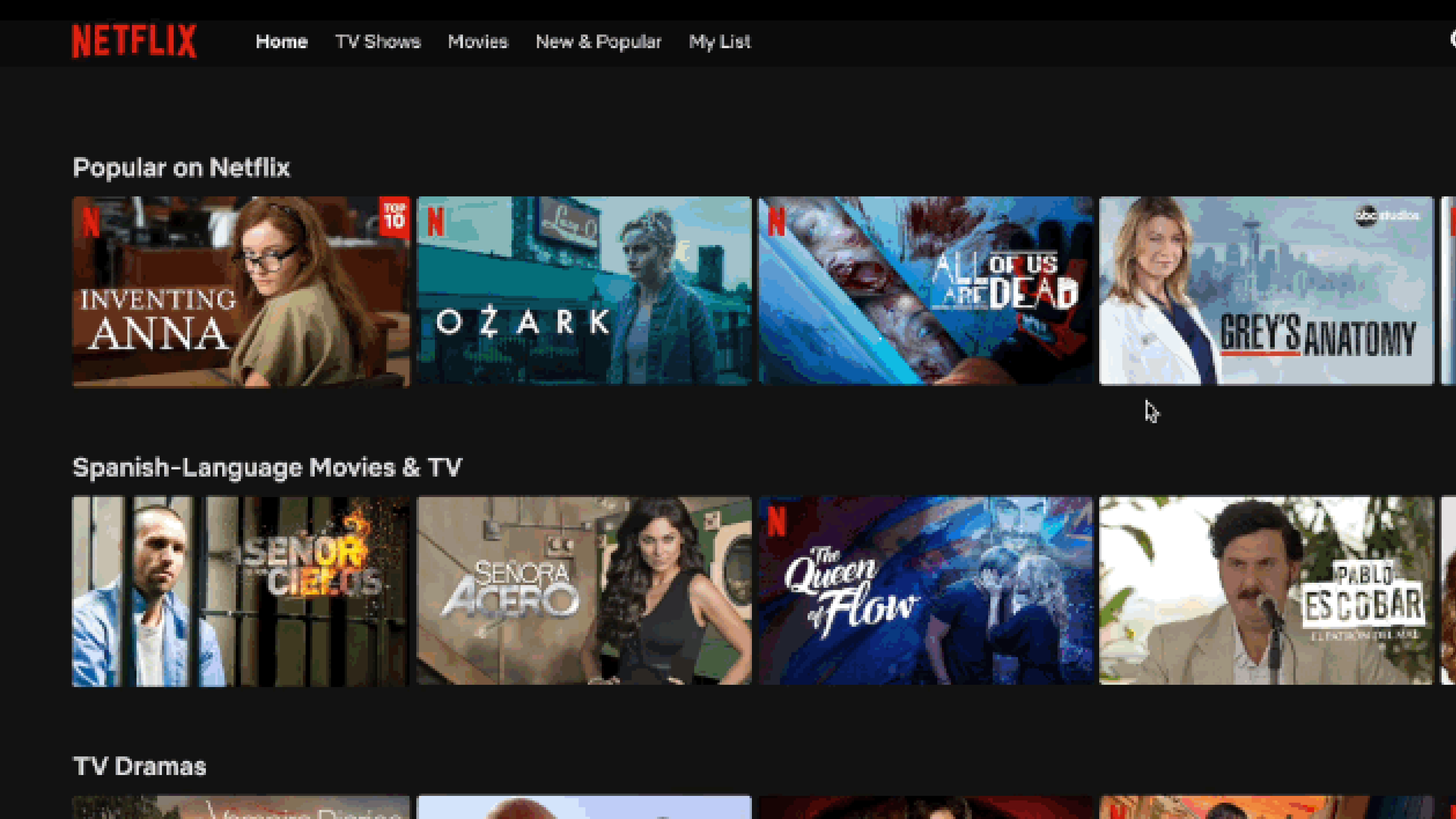
CITRIS AND THE BANATAO INSTITUTE

CITRIS POLICY LAB

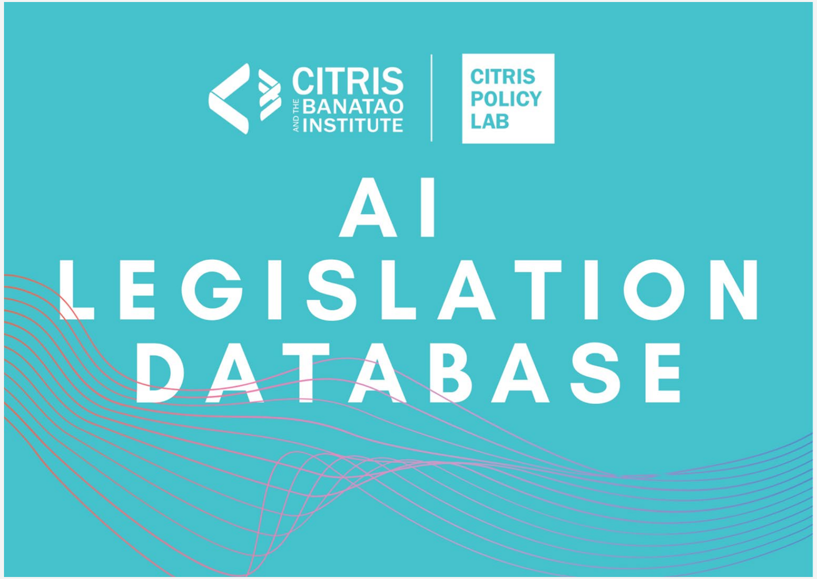


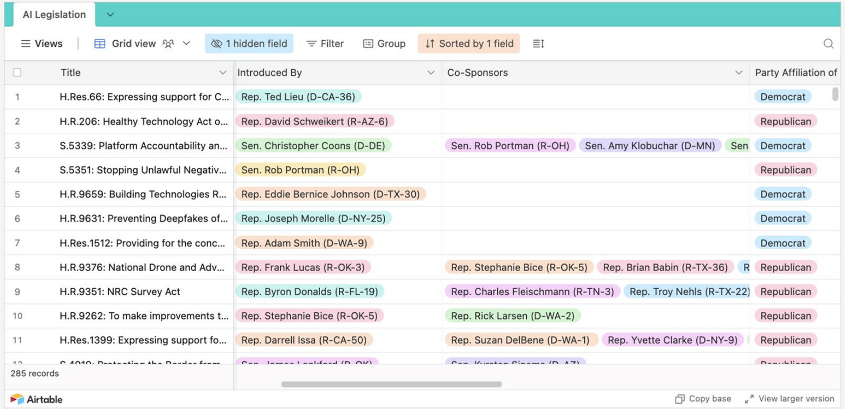


WHAT IS AI?


NETFLIX
Home
TV Shows
Movies
New & Popular
My List
Popular on Netflix
TOP 10
INVENTING ANNA
OZARK
ALL OF US ARE DEAD
abc studios
GREY'S ANATOMY
Spanish-Language Movies & TV
SEÑOR DE LOS CIELOS
SEÑORA ACERO
The Queen of Flow
PABLO ESCOBAR
EL PATRÓN DEL MAL
TV Dramas

AI LEGISLATION DATABASE (FEDERAL & CA)





	Title	Introduced By	Co-Sponsors	Party Affiliation of
1	H.Res.66: Expressing support for C...	Rep. Ted Lieu (D-CA-36)		Democrat
2	H.R.206: Healthy Technology Act o...	Rep. David Schweikert (R-AZ-6)		Republican
3	S.5339: Platform Accountability an...	Sen. Christopher Coons (D-DE)	Sen. Rob Portman (R-OH) Sen. Amy Klobuchar (D-MN) Sen	Democrat
4	S.5351: Stopping Unlawful Negativ...	Sen. Rob Portman (R-OH)		Republican
5	H.R.9659: Building Technologies R...	Rep. Eddie Bernice Johnson (D-TX-30)		Democrat
6	H.R.9631: Preventing Deepfakes of...	Rep. Joseph Morelle (D-NY-25)		Democrat
7	H.Res.1512: Providing for the conc...	Rep. Adam Smith (D-WA-9)		Democrat
8	H.R.9376: National Drone and Adv...	Rep. Frank Lucas (R-OK-3)	Rep. Stephanie Bice (R-OK-5) Rep. Brian Babin (R-TX-36) R	Republican
9	H.R.9351: NRC Survey Act	Rep. Byron Donalds (R-FL-19)	Rep. Charles Fleischmann (R-TN-3) Rep. Troy Nehls (R-TX-22)	Republican
10	H.R.9262: To make improvements t...	Rep. Stephanie Bice (R-OK-5)	Rep. Rick Larsen (D-WA-2)	Republican
11	H.Res.1399: Expressing support fo...	Rep. Darrell Issa (R-CA-50)	Rep. Suzan DelBene (D-WA-1) Rep. Yvette Clarke (D-NY-9)	Republican
12	S.4919: Protecting the Border from...	Sen. James Lankford (R-OK)	Sen. Kyrsten Sinema (D-AZ)	Republican

285 records

CITRISPolicyLab.org/AILegislation

AI DEFINED BY LAWS & INSTITUTIONS

National AI Initiative Act of 2020

AI is “a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments.”

NIST AI Risk Management Framework

An AI system is an “engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments (based off of OECD recommendation on AI: 2019; ISO/IEC 22989:2022)

AI DEFINED BY LAWS & INSTITUTIONS

EU AI Act (Article 3)

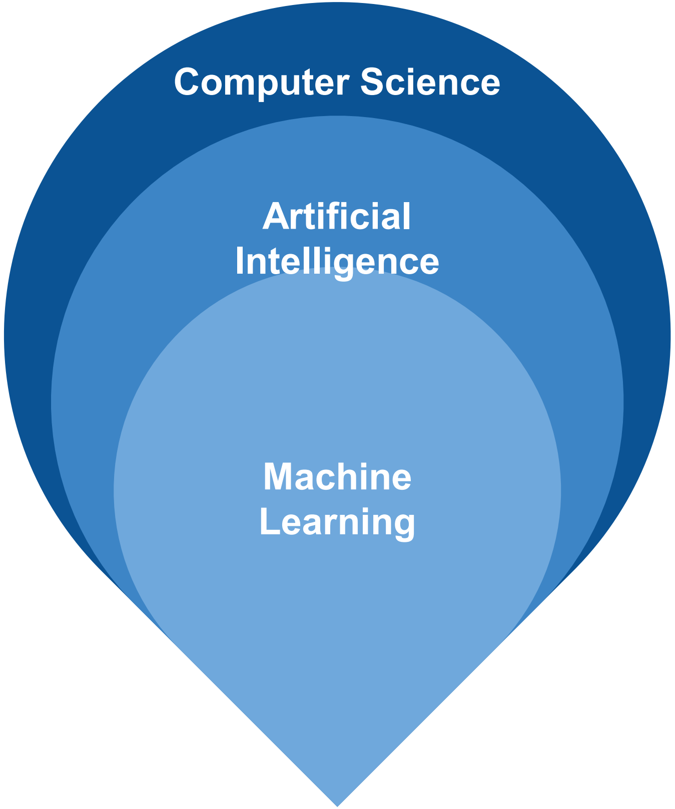
An AI system means a system that is designed to operate with elements of autonomy and that, based on machine and/or human-provided data and inputs, infers how to achieve a given set of objectives using machine learning and/or logic- and knowledge based approaches, and produces system-generated outputs such as content (generative AI systems), predictions, recommendations or decisions, influencing the environments with which the AI system interacts

AI DEFINED BY COMPUTER SCIENCE

AI refers to the ability of machines to respond to stimulation and make decisions that normally require a human level of expertise (Shubhendu & Vijay, 2013).

Machine learning (ML), the most commonly used form of AI, refers to a broad set of techniques that use data to create algorithms that are often used to predict outcomes.

- Supervised vs. Unsupervised ML
- Deep Learning
- Reinforcement Learning



MACHINE LEARNING

Supervised Machine Learning

Unsupervised Machine Learning

Reinforcement Learning

Deep Learning

Generative AI

Foundation Models

General-Purpose AI

MACHINE LEARNING

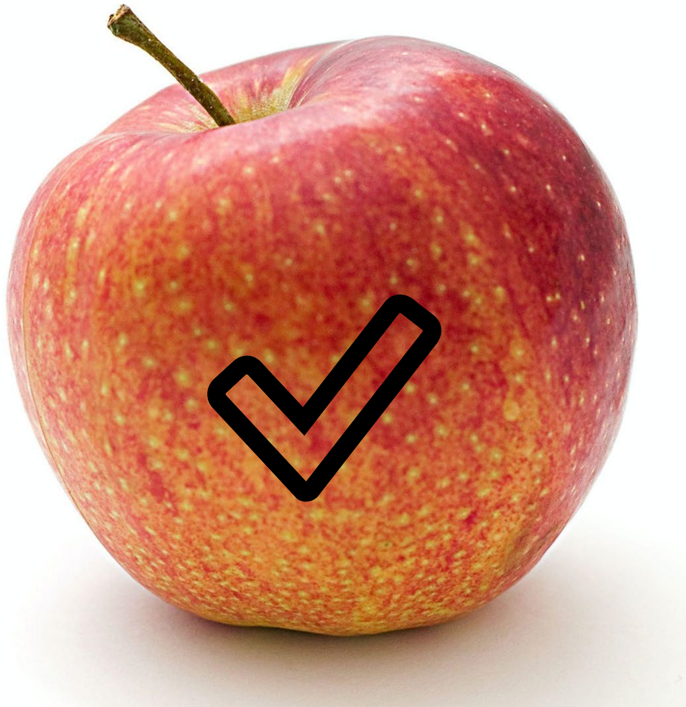
Statistical pattern recognition or correlations in data

1. **Supervised Machine Learning**
    - Labeled datasets used to train algorithms that analyze and cluster data or predict outcomes.
2. **Unsupervised Machine Learning**
    - Algorithms analyze and cluster unlabeled datasets, discover patterns.
3. **Reinforcement Learning**
    - Algorithms that learn through trial and error using feedback from its actions

ROUND

STEM

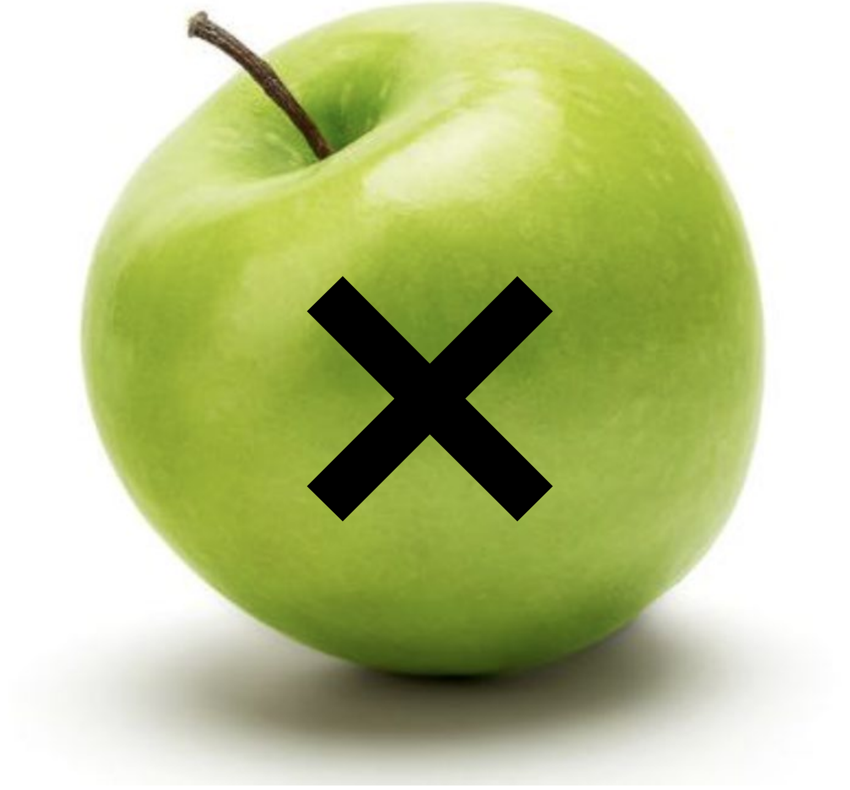
RED



ROUND ✓

STEM ✓

RED ✗

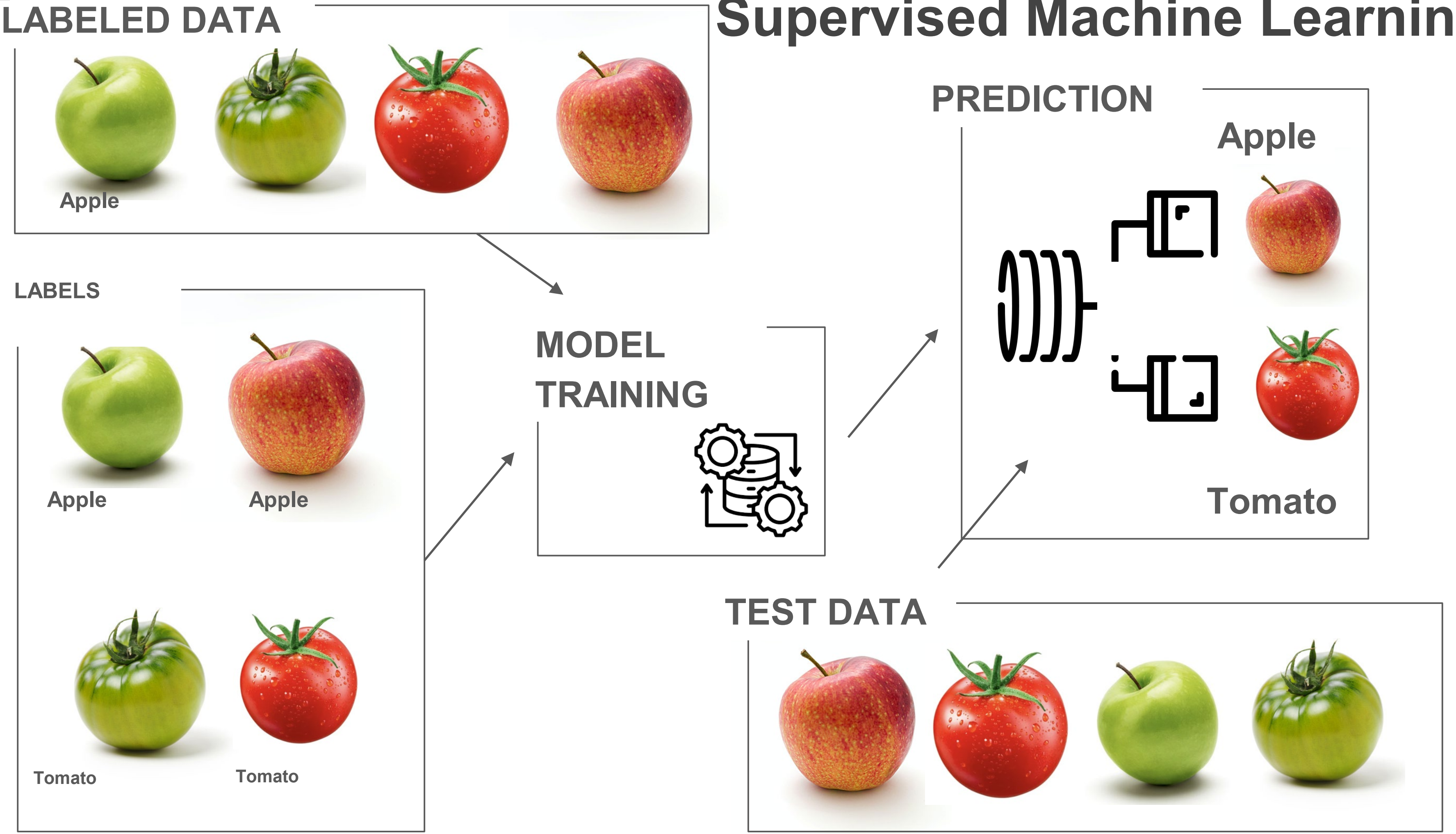


ROUND ✓

STEM ✓

RED ✓

Supervised Machine Learning



LABELED DATA

Apple

LABELS

Apple

Apple

Tomato

Tomato

MODEL TRAINING

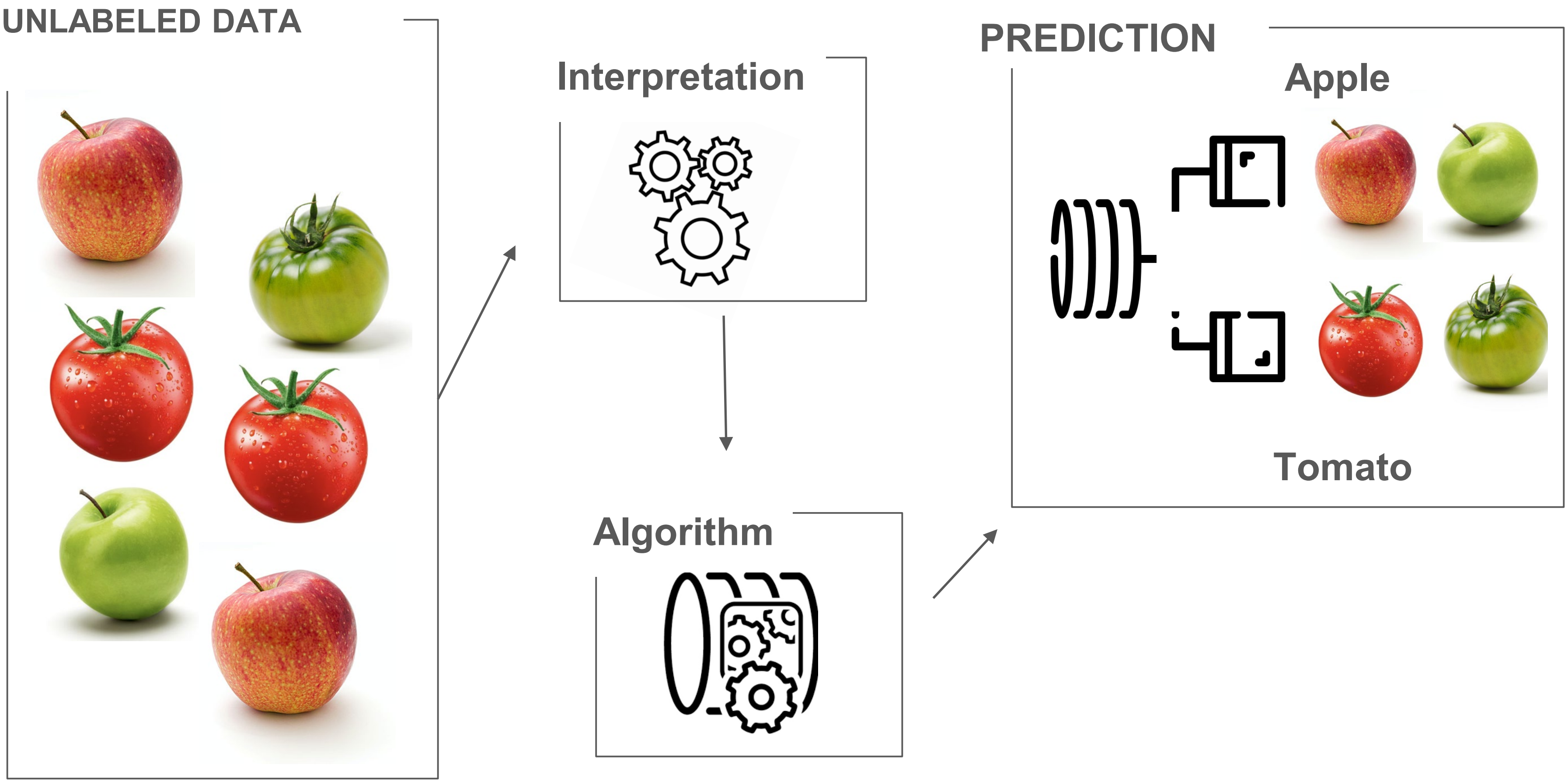
PREDICTION

Apple

Tomato

TEST DATA

Unsupervised Machine Learning



UNLABELED DATA

Interpretation

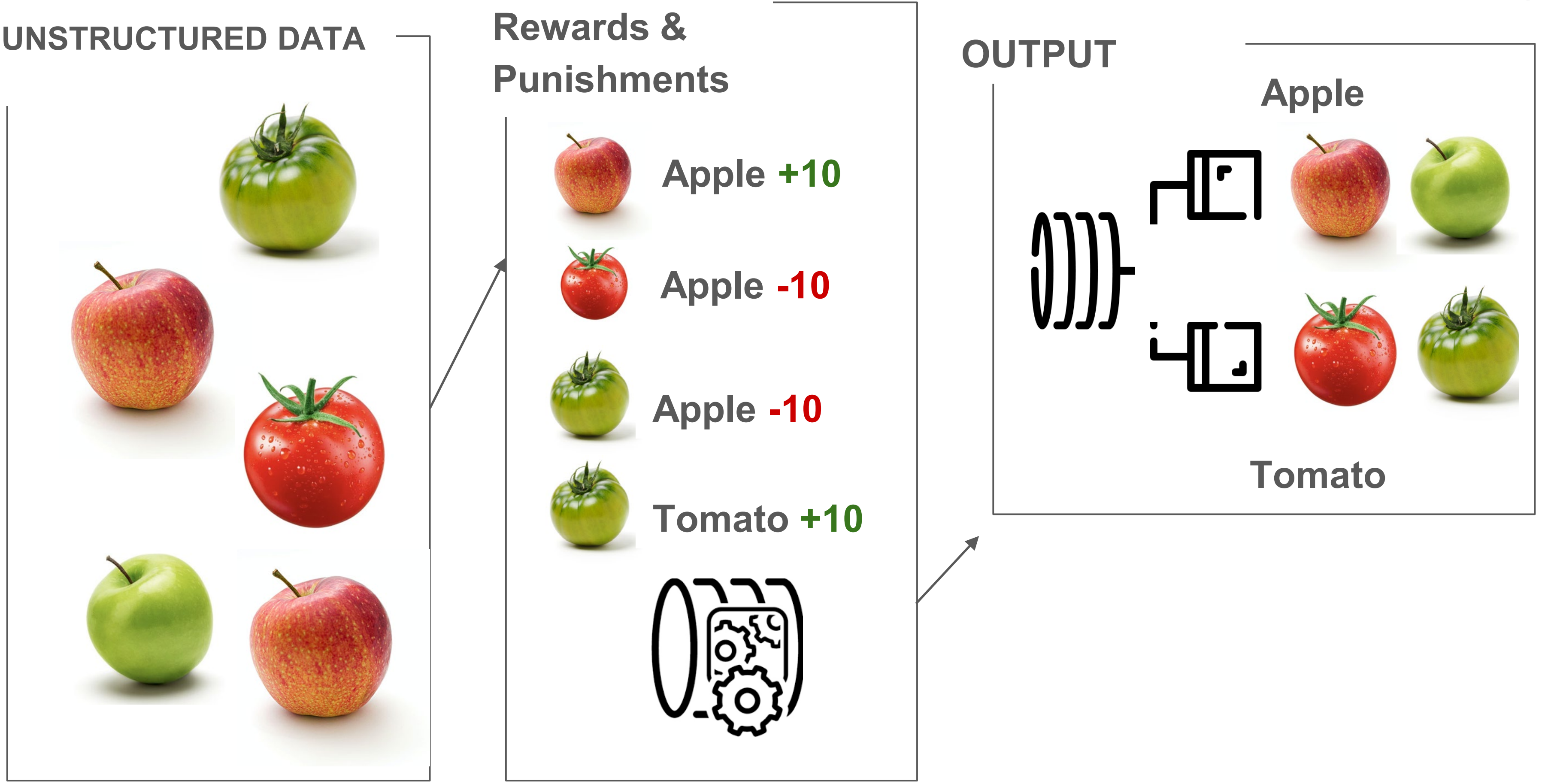
Algorithm

PREDICTION

Apple

Tomato

Reinforcement Learning



UNSTRUCTURED DATA

Rewards & Punishments

Apple +10

Apple -10

Apple -10

Tomato +10

OUTPUT

Apple

Tomato

CHALLENGES: MACHINE LEARNING

1. Supervised Machine Learning

- Can require certain levels of expertise to structure accurately
- Training supervised learning models can be very time intensive
- Datasets can have a higher likelihood of human error, resulting in algorithms learning incorrectly

2. Unsupervised Machine Learning

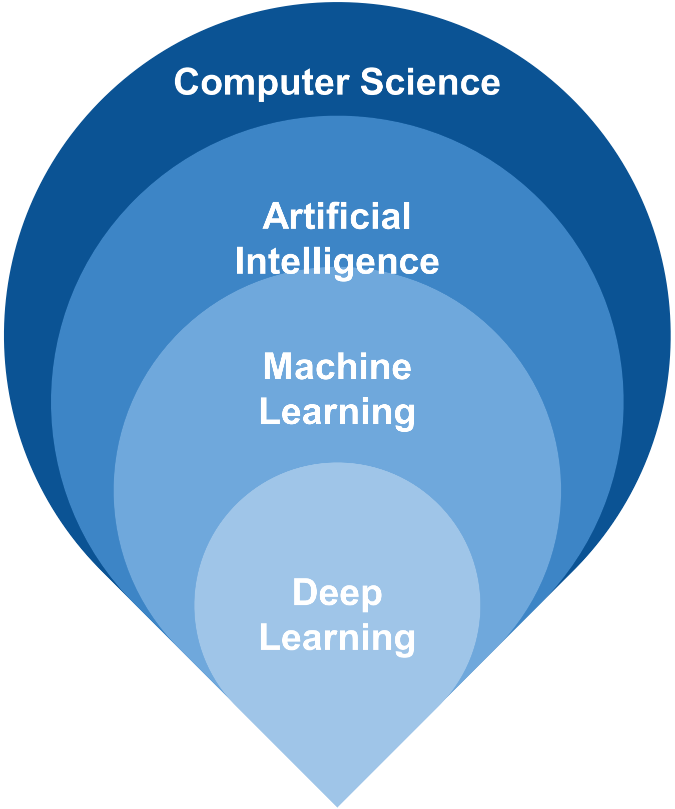
- Computational complexity due to a high volume of training data
- Higher risk of inaccurate results
- Lack of transparency into the basis on which data were clustered

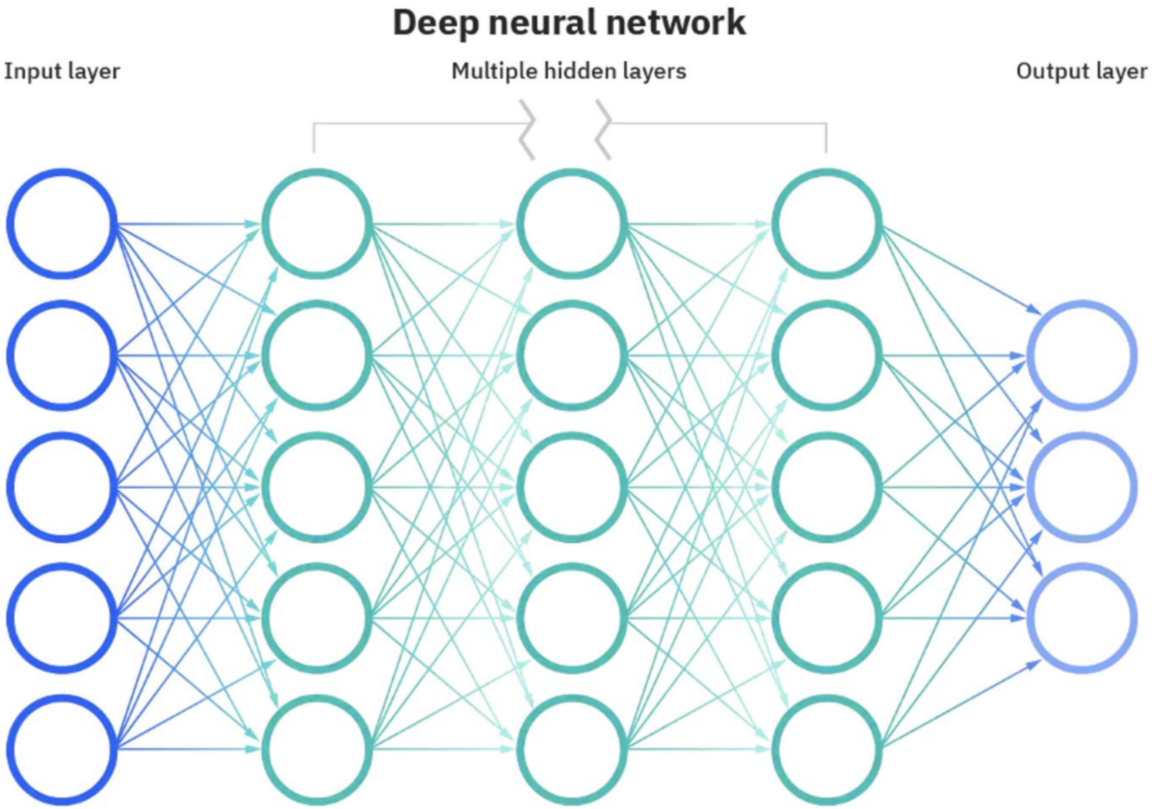
3. Reinforcement Learning

- All of the Above &...
- Faulty reward functions create unintended behaviors

DEEP LEARNING

- Concept around since 1950s (Frank Rosenblatt)
- A subset of machine learning
- More complex
- Mimics the human brain (i.e., how neurons fire in brain)
- Ingest & process unstructured data
- Automates feature extraction (e.g., dog ears vs. cat ears)
- Classify and cluster data





Source: https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/

CHALLENGES: DEEP LEARNING

- Large amounts of data
- Powerful computing
- Lack of transparency
- Faulty reward functions create unintended behaviors

GENERATIVE AI

Deep learning models that can generate high-quality text, images, audio, and other content based on the data they were trained on.

Midjourney

CHATGPT

Bard

FOUNDATION MODELS

AI systems with broad capabilities that can be adapted to a range of different, more specific purposes.

The original model provides a “foundation” on which other things are built

The large language model GPT-4 is the foundation model of ChatGPT

AI GOVERNANCE

US Federal AI Landscape

2019

- United States adopts OECD Principles on Artificial Intelligence
- Executive Order “Maintaining American Leadership in AI” (2019)

2020

- AI in Government Act of 2020
- Executive Order “Promoting the Use of Trustworthy AI in the Federal Government (2020)

2021

- National AI Initiative Act of 2020 (became law in January 2021)
  - National AI Initiative Office (housed within White House OSTP)

2022

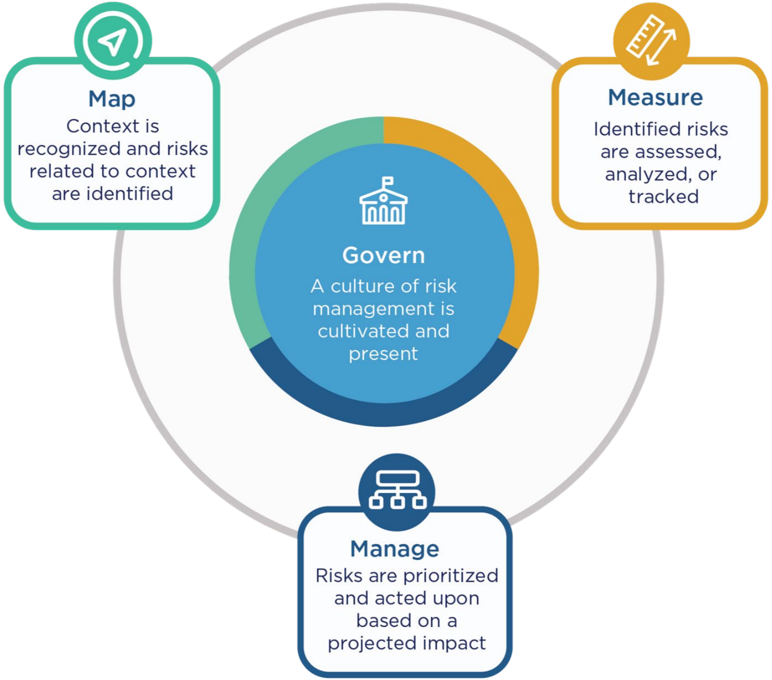
  - National AI Advisory Committee

2023

- NIST AI Risk Management Framework
- AI Bill of Rights
- White House Voluntary AI Commitments
- Sens. Blumenthal & Hawley introduce framework to guide AI governance and subsequent bills
- Sen. Schumer’s AI Summit & “Safe Innovation Framework for AI Policy”
- White House Executive Order on AI

NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

AI Risk Management Framework

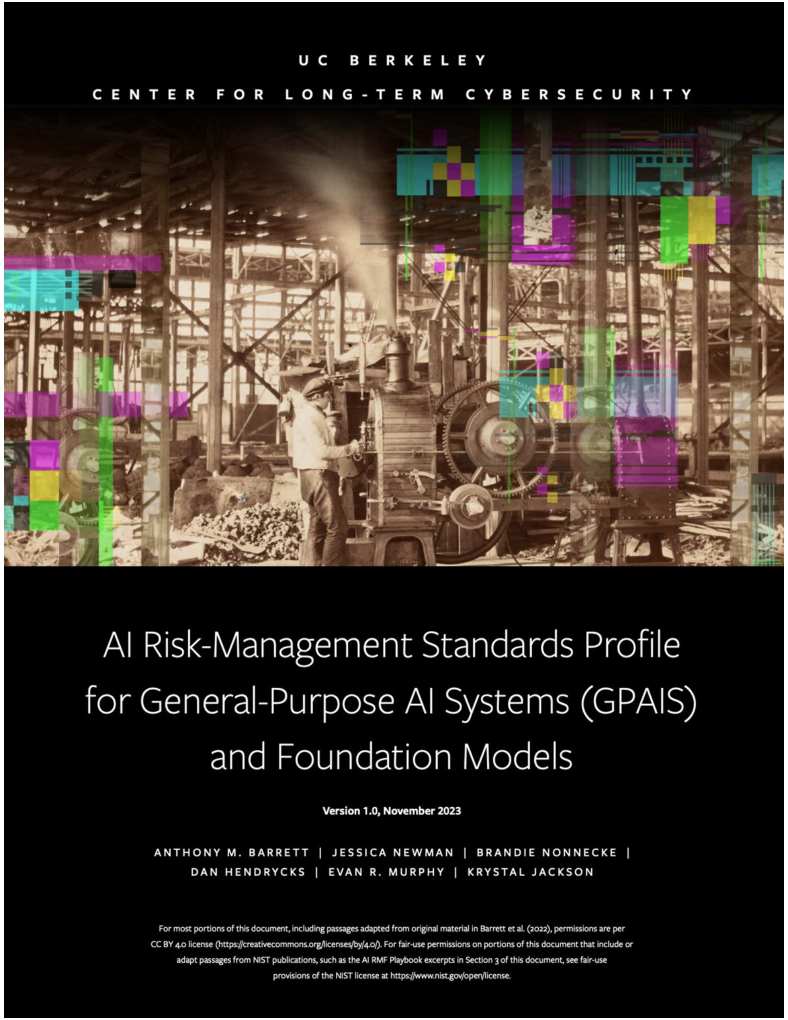


Map
Context is recognized and risks related to context are identified

Measure
Identified risks are assessed, analyzed, or tracked

Govern
A culture of risk management is cultivated and present

Manage
Risks are prioritized and acted upon based on a projected impact



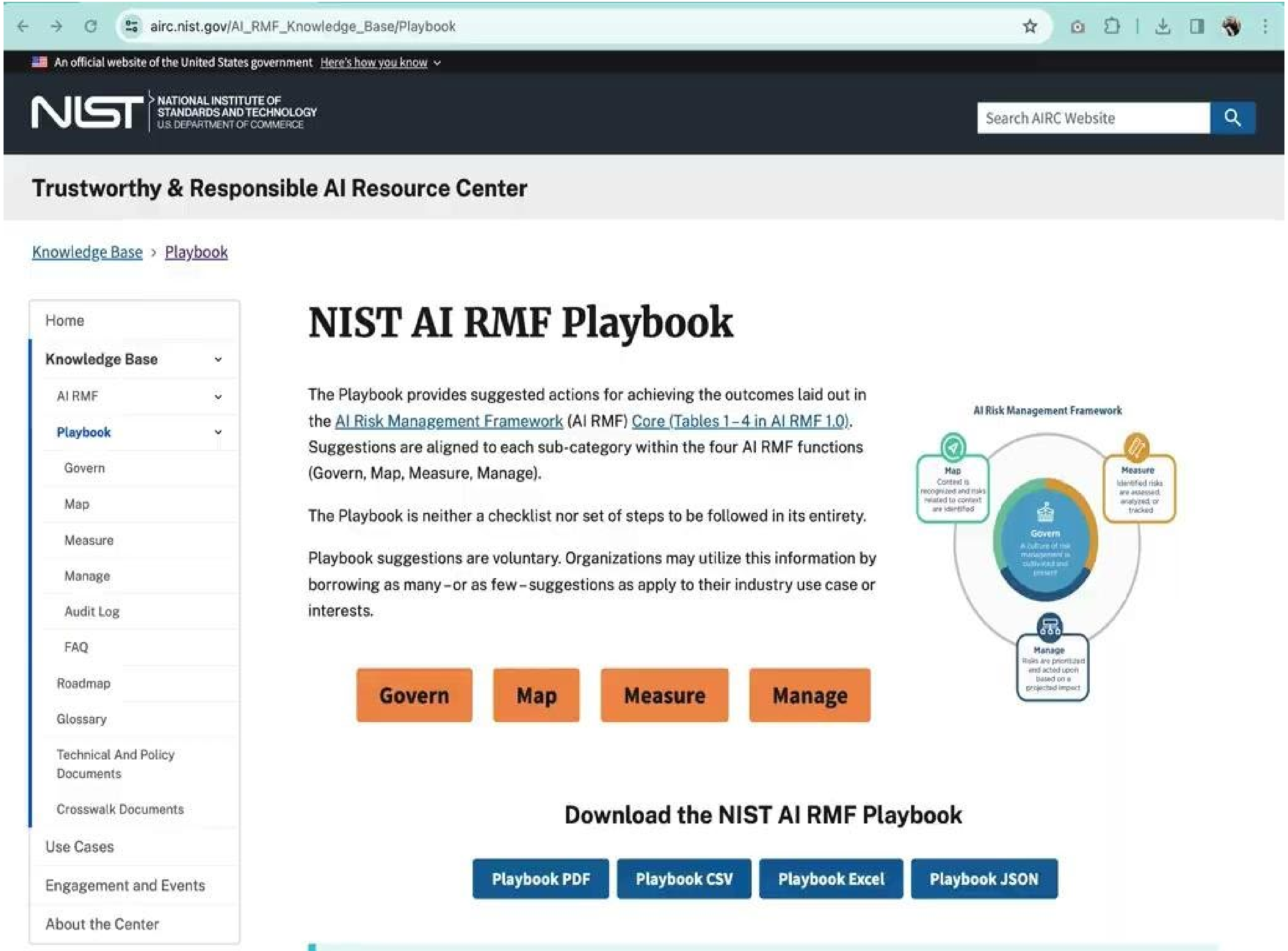
UC BERKELEY

CENTER FOR LONG-TERM CYBERSECURITY

AI Risk-Management Standards Profile for General-Purpose AI Systems (GPAIS) and Foundation Models

Version 1.0, November 2023

ANTHONY M. BARRETT | JESSICA NEWMAN | BRANDIE NONNECKE |
DAN HENDRYCKS | EVAN R. MURPHY | KRYSTAL JACKSON



airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

An official website of the United States government
Here's how you know
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE
Search AIRC Website

Trustworthy & Responsible AI Resource Center

Knowledge Base › Playbook

NIST AI RMF Playbook

The Playbook provides suggested actions for achieving the outcomes laid out in the AI Risk Management Framework (AI RMF) Core (Tables 1 – 4 in AI RMF 1.0). Suggestions are aligned to each sub-category within the four AI RMF functions (Govern, Map, Measure, Manage).

The Playbook is neither a checklist nor set of steps to be followed in its entirety.

Playbook suggestions are voluntary. Organizations may utilize this information by borrowing as many – or as few – suggestions as apply to their industry use case or interests.

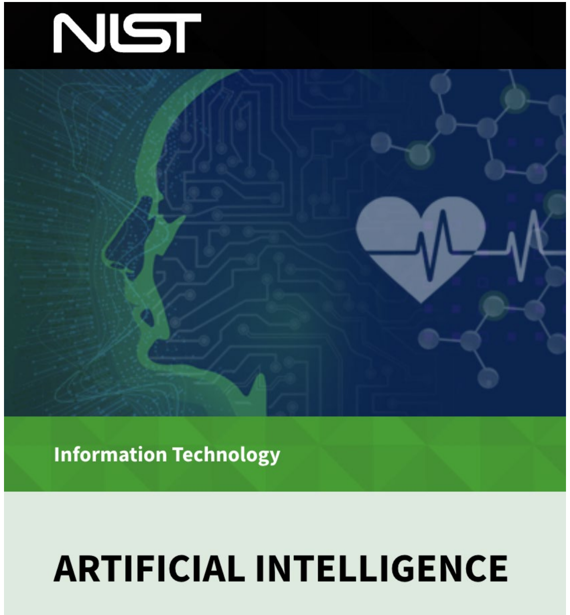
AI Risk Management Framework
Map
Context is recognized and risks related to context are identified
Measure
Identified risks are assessed, analyzed, or tracked
Govern
A culture of risk management is cultivated and present
Manage
Risks are prioritized and acted upon based on a projected impact
Govern
Map
Measure
Manage

Download the NIST AI RMF Playbook

European Union

- EU AI Act (passed)
  - Most comprehensive AI legislation globally
  - Puts in place requirements on high-risk AI systems
- Digital Services Act (passed)
- Digital Markets Act (passed)
- Data Governance Act (passed)
- EU General Data Protection Regulation (passed)
  - Article 22 “The data subject shall have the right not to be subject to **a decision** based **solely** on **automated processing,** including **profiling**, which produces **legal effects** concerning him or her or similarly significantly affects him or her.”

AI Standards & Guidelines



ISO/IEC JTC 1/SC 42

Artificial intelligence

IEEE ETHICS IN ACTION

in Autonomous and Intelligent Systems

OCEANIS

OPEN COMMUNITY FOR ETHICS IN AUTONOMOUS AND INTELLIGENT SYSTEMS

The Global AI Standards Repository

Third-party Auditors, Evaluators, Licensors, Certifiers

Auditors

ORCAA

Parity AI

Evaluators

Credo.ai

ARC Evals

Licensors

Responsible AI Licenses (RAIL)

Certifiers

Responsible AI Institute

CONTACT

Brandie Nonnecke, PhD
Director, CITRIS Policy Lab
Assoc. Research Professor, Goldman School of Public Policy
nonnecke@berkeley.edu | @BNonnecke

GLOSSARY

AI Bias - Computational or statistical bias is a systematic error or deviation from the true value of a prediction that originates from a model's assumptions or the data itself. Human or cognitive bias refers to inaccurate individual judgment or distorted thinking, while systemic bias leads to systemic prejudice, favoritism, and/or discrimination in favor of or against an individual or group. Bias can impact outcomes and pose a risk to individual rights and liberties (NIST, 2022; IAPP, 2023)

AI Risks - Like risks for other types of technology, AI risks can emerge in a variety of ways and can be characterized as long- or short-term, high- or low-probability, systemic or localized, and high- or low-impact (NIST AI RMF, 2023)

AI Fairness - An attribute of an AI system that ensures equal and unbiased treatment of individuals or groups in its decisions and actions in a consistent, accurate manner. It means the AI system's decisions should not be affected by certain sensitive attributes like race, gender or religion (IAPP, 2023)

Trustworthy AI - Often used interchangeably with the terms responsible AI and ethical AI, which all refer to principle-based AI development and governance, including the principles of security, safety, transparency, explainability, accountability, privacy, nondiscrimination/non-bias, among others (IAPP, 2023)