# 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**

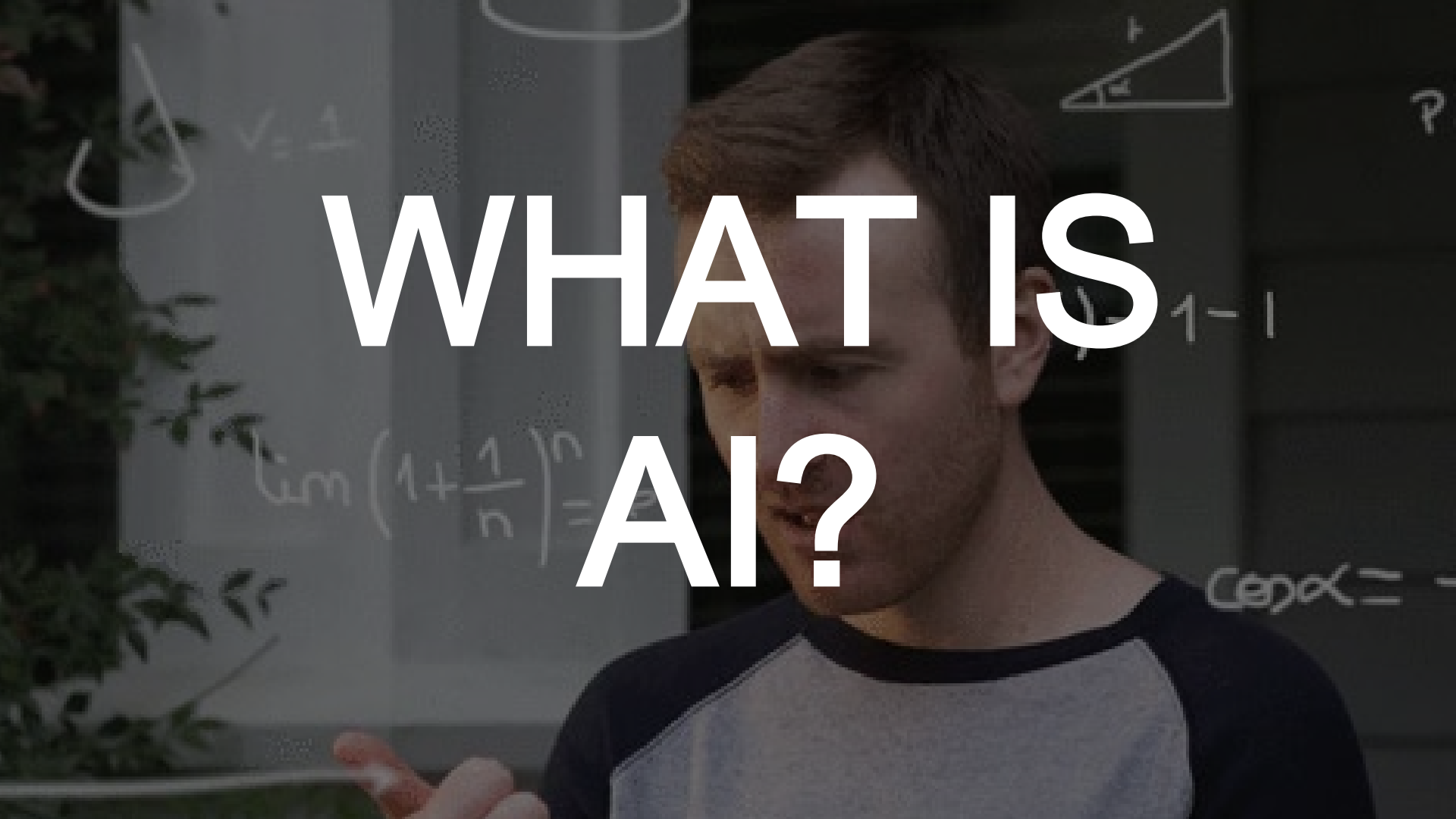
**2024**


CITRIS
AND THE BANATAO INSTITUTE
CITRIS
POLICY
LAB

Any sufficiently advanced technology is indistinguishable from

# MAGIC

- Arthur C. Clark, Author,
2001: A Space Odyssey


WHAT IS
AI?


TECHYPE
Debunking Emerging Tech
WITH
BRANDIE NONNECKE, PHD
LEARN MORE > TECHYPE.ORG
CITRIS
BANATAO INSTITUTE
CITRIS POLICY LAB
Berkeley Public Policy
The Goldman School

NETFLIX Home TV Shows Movies New & Popular My List

## Popular on Netflix





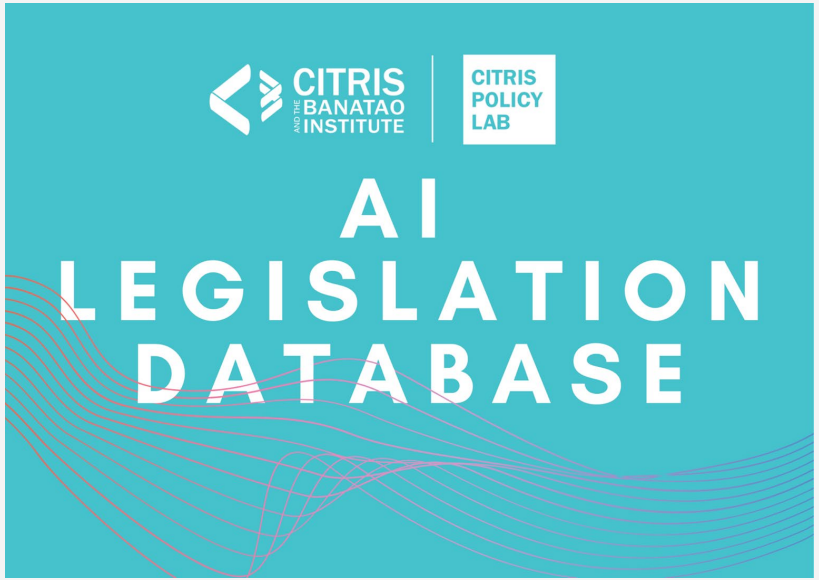
## Spanish-Language Movies & TV

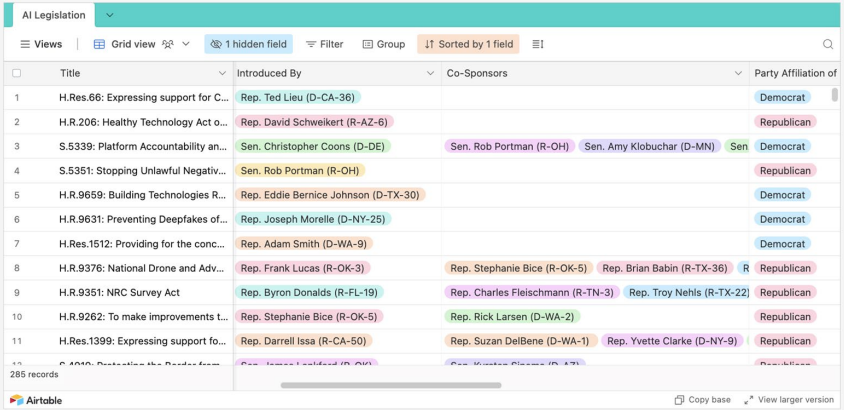




## TV Dramas

# AI LEGISLATION DATABASE (FEDERAL & CA)




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)**

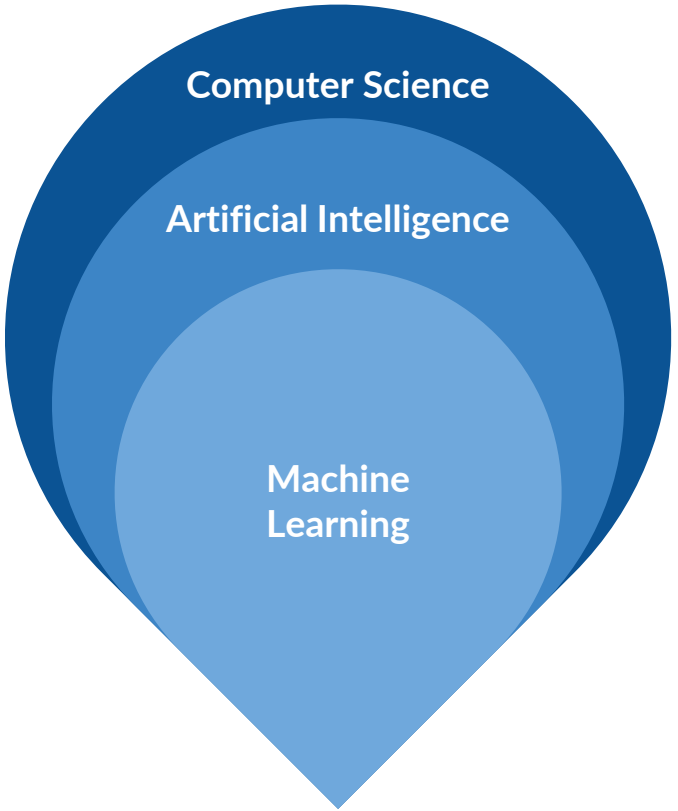
An AI system means a system that is designed to operate with elements of autonomy and that, based on machine and/or human-provided data and inputs, infers how to achieve a given set of objectives using machine learning and/or logic- and knowledge based approaches, and produces system-generated outputs 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

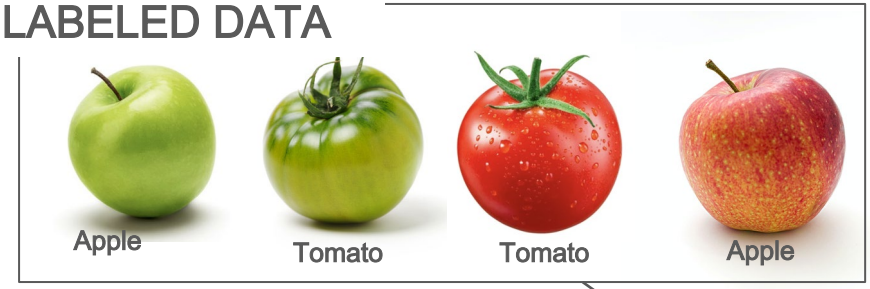
# MACHINE LEARNING

Statistical pattern recognition or correlations in data

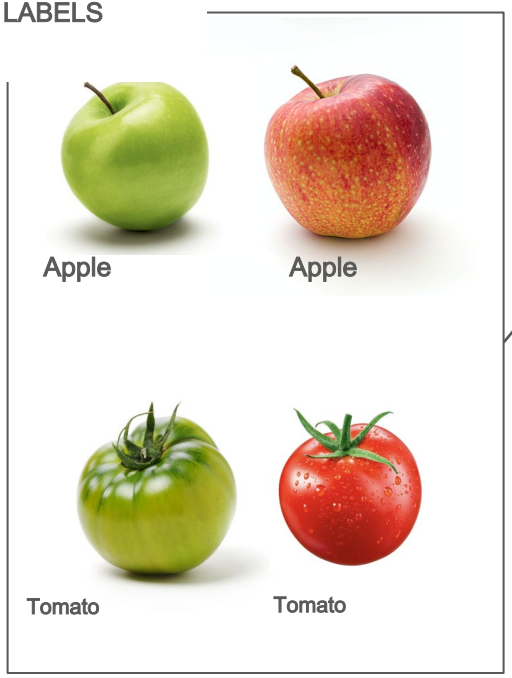
1. **Supervised Machine Learning**
   - Labeled datasets used to train algorithms that analyze and cluster data or predict outcomes.
2. **Unsupervised Machine Learning**
   - Algorithms analyze and cluster unlabeled datasets, discover patterns.
3. **Reinforcement Learning**
   - Algorithms that learn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

# Supervised Machine Learning

LABELED DATA

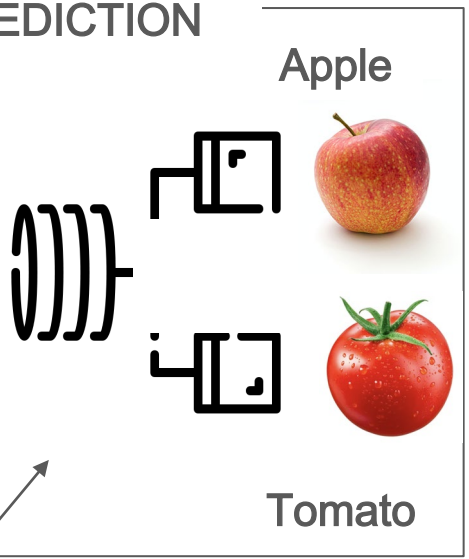



LABELS

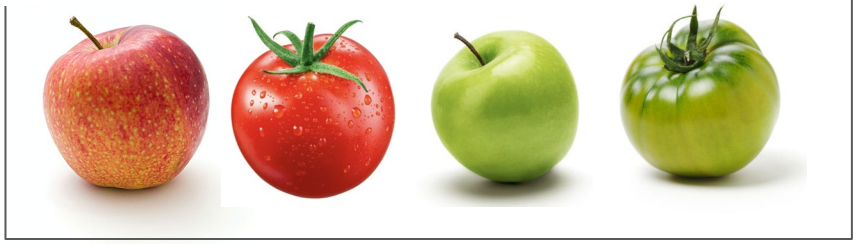



MODEL TRAINING

PREDICTION




TEST DATA



# Unsupervised Machine Learning

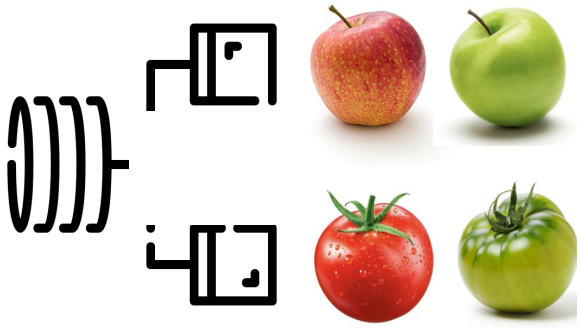
UNLABELED DATA

Interpretation

Algorithm

PREDICTION

Apple



Tomato

# Reinforcement Learning

UNSTRUCTURED DATA

Rewards & Punishments

Apple +10

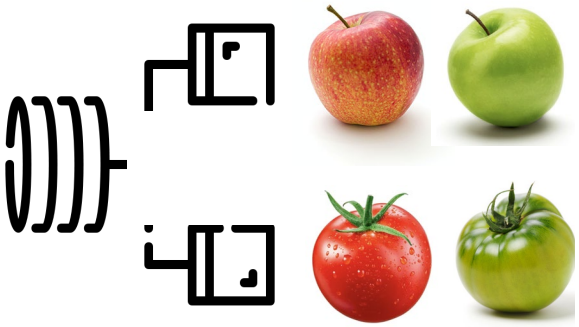
Apple -10

Apple -10

Tomato +10

OUTPUT

Apple



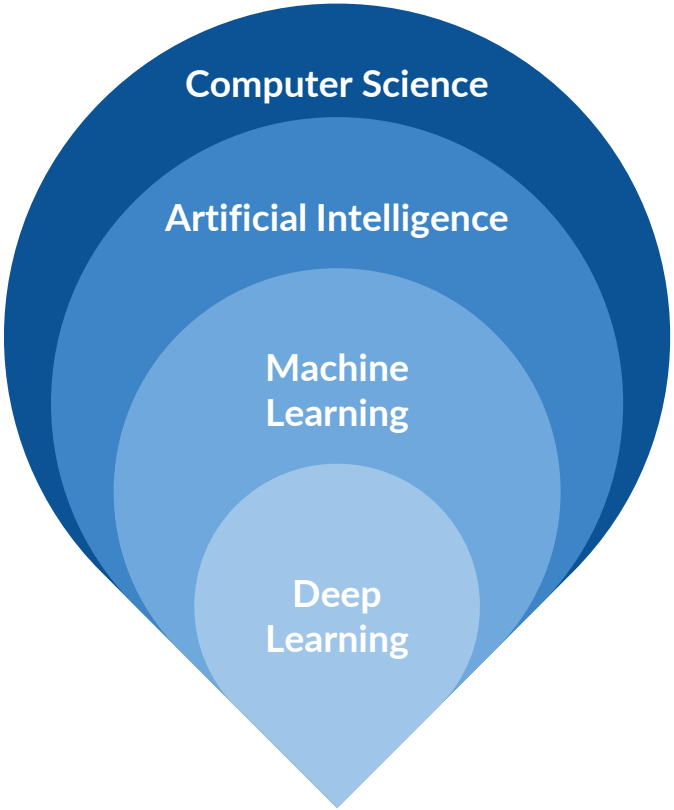
Tomato

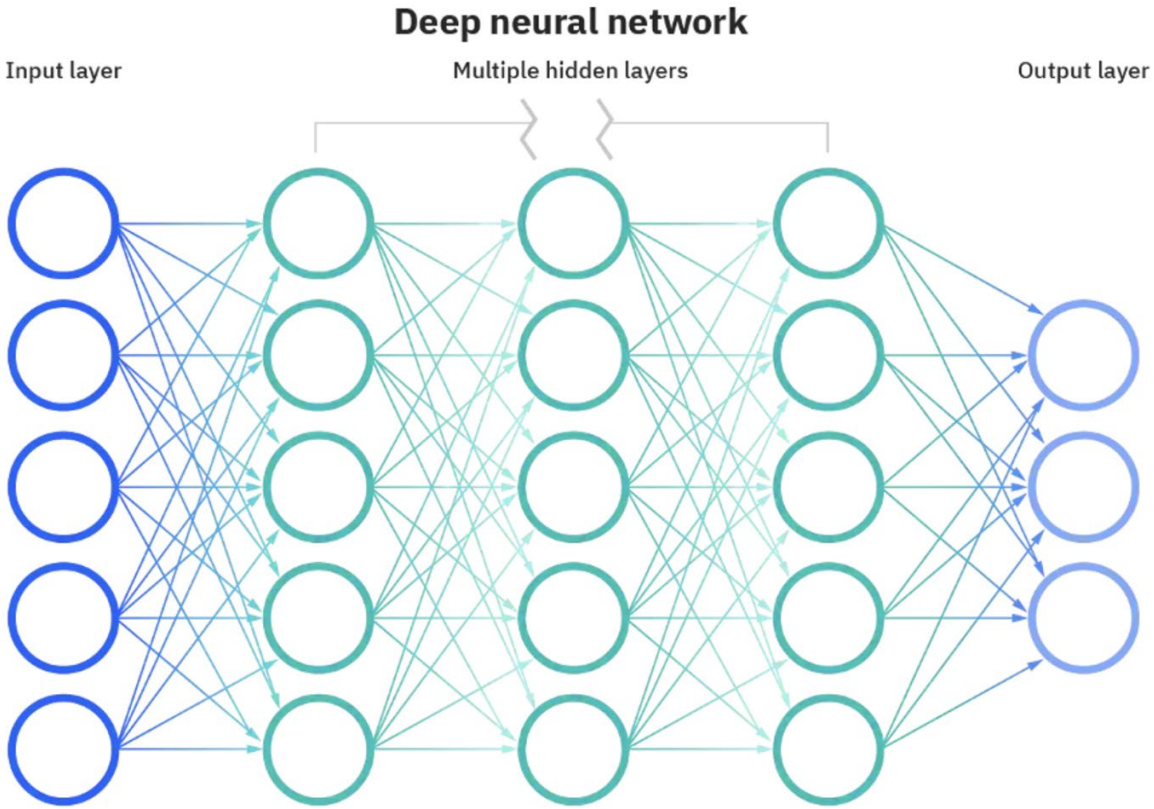
# CHALLENGES: MACHINE LEARNING

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

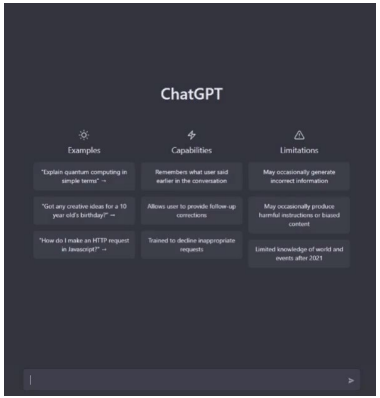
CHATGPT

# 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-3.5 is the foundation model of ChatGPT




**Deep Learning**
**Generative AI**

# LLMs → LMMs

# GENERAL-PURPOSE AI

## EU AI Act (Article 3)

An AI system that - irrespective of how it is placed on the market or put into service, including as open-source software - is intended by the provider to perform generally applicable functions such as image and speech recognition, audio and video generation, pattern detection, question answering, translation and others; a general purpose AI system may be used in a plurality of contexts and be integrated in a plurality of other AI systems

# Regulation of AI

# *Ex-ante vs. Ex-post*

# 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)

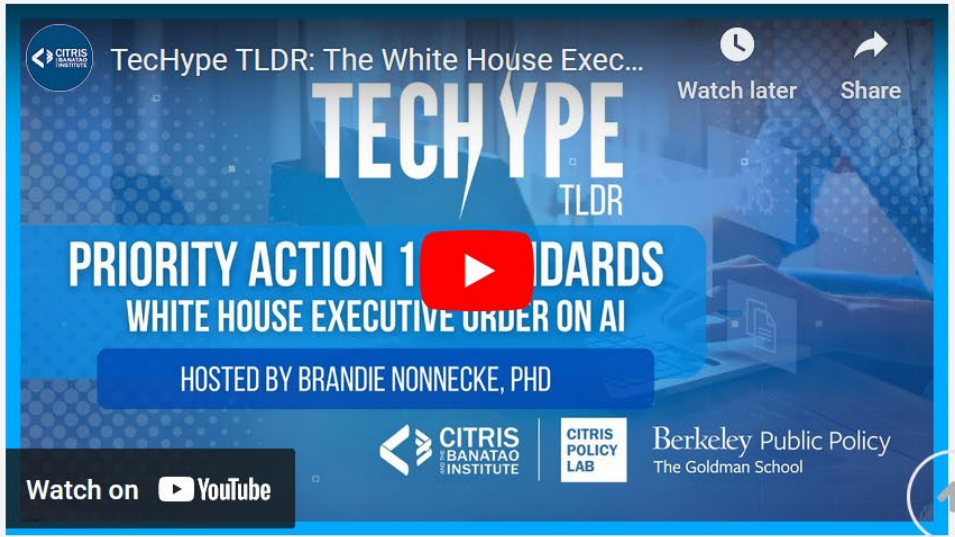
**2021**
- National AI Initiative Act of 2020 (became law in January 2021)
    - National AI Initiative Office (housed within White House OSTP)

**2022**
    - National AI Advisory Committee

# US Federal AI Landscape

- 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


TecHype TLDR: The 8 Priority Actions o...
Watch later
Share
White House
Executive Order on AI
8 PRIORITY ACTIONS
1 NEW STANDARDS FOR AI SAFETY
2 PROTECTING AMERICANS' PRIVACY
3 ADVANCING EQUITY & CIVIL RIGHTS
4 STANDING UP FOR CONSUMERS, PATIENTS & STUDENTS
5 SUPPORTING WORKERS
TECHYPE
TLDR
ORITY ACTIONS
IN THE HOUSE EXECUTIVE ORDER ON AI
HOSTED BY BRANDIE NONNECKE, PHD
CITRIS BANATAO INSTITUTE
CITRIS POLICY LAB
Berkeley Public Policy
The Goldman School
Watch on YouTube


TecHype TLDR: The White House Exec...
Watch later
Share
TECHYPE
TLDR
PRIORITY ACTION 1 DARDS
WHITE HOUSE EXECUTIVE ORDER ON AI
HOSTED BY BRANDIE NONNECKE, PHD
CITRIS BANATAO INSTITUTE
CITRIS POLICY LAB
Berkeley Public Policy
The Goldman School
Watch on YouTube

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

## AI Risk Management Framework

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.


**Govern** **Map** **Measure** **Manage**

## Download the NIST AI RMF Playbook

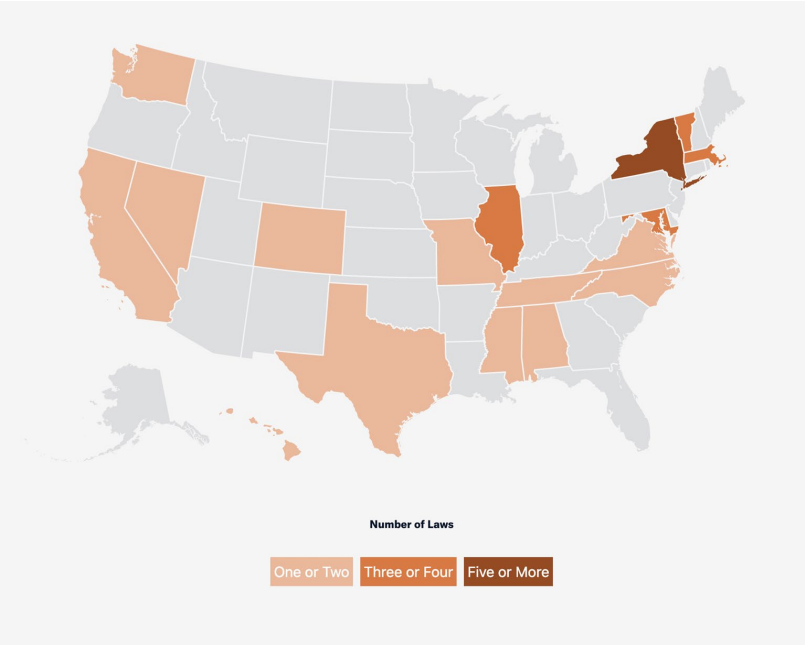
**Playbook PDF** **Playbook CSV** **Playbook Excel** **Playbook JSON**

# Federal AI Landscape

Federal agencies and departments play a central role in developing and overseeing AI

- National Institute for Standards and Technology (NIST)
- Federal Trade Commission (FTC)
- Department of Commerce (DOC)
- National Science Foundation (NSF)
- Department of Energy (DOE)
- Food and Drug Administration (FDA)
- Department of Defense (DOD)
- *And many more…*

# States AI Landscape




U.S. Chamber of Commerce, 2022

**California** - Anti-Deepfake Law, Bots Bill, AB-331(proposed), EO on GenAI

**Illinois** - Biometric Information Privacy Act (BIPA)

**New York** - Commission on the Future of Work & NYC law on AI-enabled tools for HR

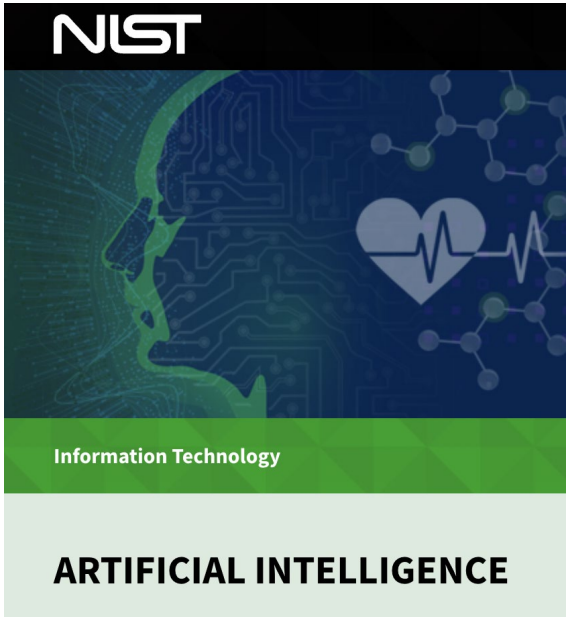
***Key Themes:***

- Advisory boards/councils/task forces
- Banning biometrics/facial recognition
- Workforce development
- Privacy
- Non-discrimination & Auditing

# European Union

- EU AI Act (passed, final text to be released in early 2024)
  - 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

# Intergovernmental & Multistakeholder Initiatives

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

Launched in June 2020 with 15 members, GPAI is the fruition of an idea developed within the G7.

Today, GPAI's 25 members are Australia, Belgium, Brazil, Canada, Czech Republic, Denmark, France, Germany, India, Ireland, Israel, Italy, Japan, Mexico, the Netherlands, New Zealand, Poland, the Republic of Korea, Singapore, Slovenia, Spain, Sweden, the United Kingdom, the United States and the European Union.

OECD.AI Policy Observatory

## OECD Network of Experts on AI (ONE AI)

The OECD Network of Experts on AI (ONE AI) provides policy, technical and business expert input to inform OECD analysis and recommendations. It is a multi-disciplinary and multi-stakeholder group.


PARTNERSHIP ON AI

# Industry Practices & Policies

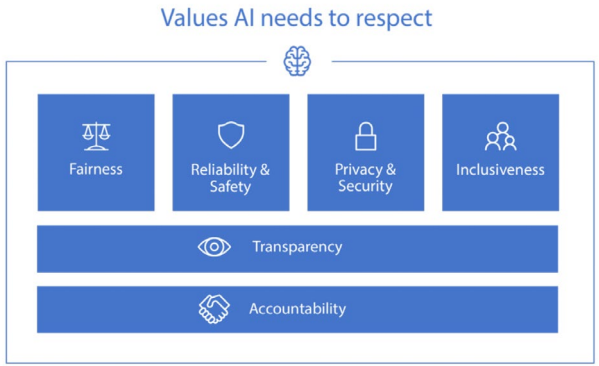
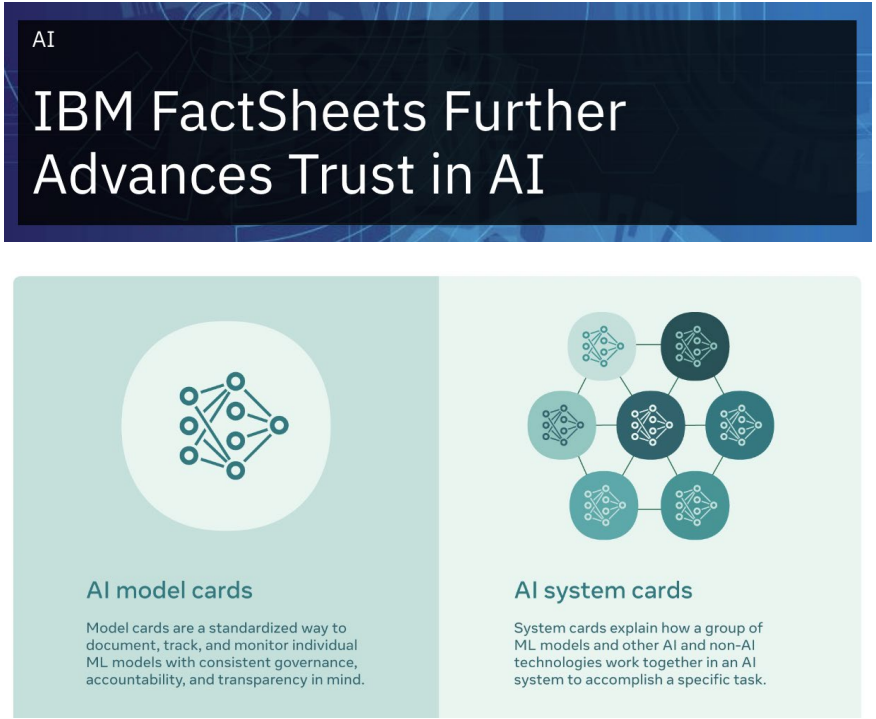




Chart 5.
Source: Microsoft Corporation



# 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


TECHYPE
Debunking Emerging Tech
WITH
BRANDIE NONNECKE, PHD
LEARN MORE > TECHYPE.ORG
CITRIS
BANATAO INSTITUTE
CITRIS POLICY LAB
Berkeley Public Policy
The Goldman School


# TecHype: Demystifying AI & Other Emerging Tech

TecHype

Technology

Listen on Apple Podcasts ↗ Requires macOS 11.4 or higher

Join us for a captivating preview of TecHype's first season featuring top experts who unravel the mysteries of AI, social media, deepfakes, and more. Episodes will be released weekly.

Our first season features enlightening discussions with AI pioneers, including UC Berkeley Prof. Stuart Russell on the responsible development and use of AI and Prof. Hany Farid on the rise of deepfakes and strategies to safeguard democracy.

Award-winning NY Times Journalist Mike Isaac guides listeners through the AI development debate raging in Silicon Valley. World-renowned actor-director Alex Winter provides an insider's perspective on how AI is transforming the entertainment industry. Yoel Roth, Twitter's former head of Trust & Safety, provides unparalleled insight into content moderation practices and Prof. Joan Donovan reveals the evolving tactics nefarious actors are using to spread disinformation online.

TecHype also produces TLDR (Too long; Didn't read) shorts that analyze and summarize emerging tech policies, regulations, and laws.

TecHype is more than a show; it's a movement towards


1 episode

TecHype provides an eye-opening journey through our modern digital world, challenging perceptions and exploring the fine line more

## TecHype

Brandie Nonnecke, PhD

Technology

Listen on Apple Podcasts ↗ Requires macOS 11.4 or higher

JAN 12, 2024

**TecHype: Demystifying AI & Other Emerging Tech** ›

Join us for a captivating preview of TecHype's first season featuring top experts who unravel the mysteries of AI, social media, deepfakes, and more. Episodes will be released weekl...

# 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)