ETHICAL & RESPONSIBLE USE OF GENERATIVE AI

GABRIELLE HARRIS

BYU
BRIGHAM YOUNG
UNIVERSITY

BYU
HAWAII

BYU
IDAHO

ENSIGN
COLLEGE

Church Educational System
Privacy Center

DISCUSSION TOPICS:

- Introductions
- Why do we care about the ethical & responsible use of AI?
- What are we doing about it at BYU?
- What do you need to do as you use GenAI?
- Additional resources

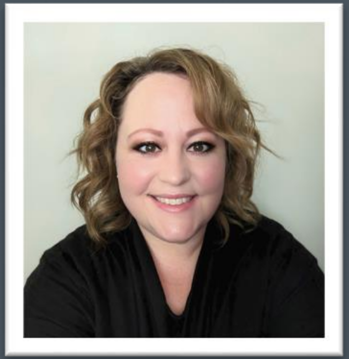
INTRODUCTIONS

—

CES PRIVACY CENTER

CES PRIVACY CENTER:

HOWARD LOOS



GABRIELLE HARRIS

BYU BRIGHAM YOUNG UNIVERSITY

BYU HAWAII

BYU IDAHO

ENSIGN COLLEGE

WHY DO WE CARE ABOUT THE ETHICAL & RESPONSIBLE USE OF AI?

CES PRIVACY CENTER

OUR INSTITUTIONAL VALUES & GOALS

ELDER DAVID A. BEDNAR

BYU Devotional January 23, 2024

"Innovations such as artificial intelligence have the potential to both

(1) assist you in receiving magnificent blessings

and

(2) diminish and suffocate your moral agency."

“While generative artificial intelligence may be quick to offer information, it can never replace revelation or generate truth. If something does not feel right or is inconsistent with what you know is true, seek to discern before believing.”

Elder Gerrit W. Gong

March 13, 2024

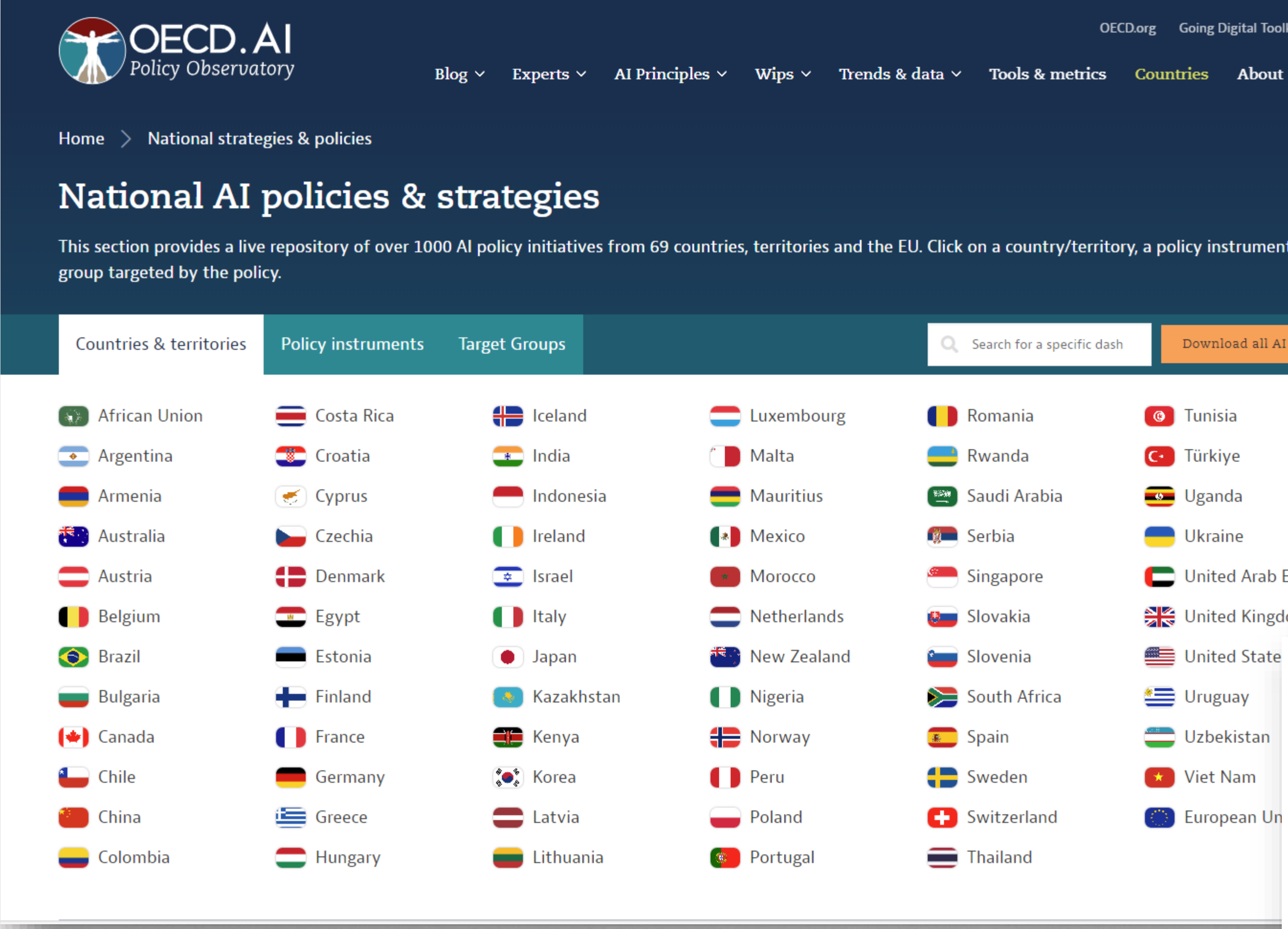
ETHICS & AI

“In no other field is the ethical compass more relevant than in artificial intelligence. These general-purpose technologies are re-shaping the way we work, interact, and live. The world is set to change at a pace not seen since the deployment of the printing press six centuries ago. AI technology brings major benefits in many areas, but without the ethical guardrails, it risks reproducing real world biases and discrimination, fueling divisions and threatening fundamental human rights and freedoms. “

-GABRIELA RAMOS

Assistant Director-General, Social and Human Sciences, UNESCO
(United Nations Educational, Scientific and Cultural Organization)

REGULATIONS & ENFORCEMENTS



OECD.AI Policy Observatory

OECD.org Going Digital Toolkit | EN

Blog Experts AI Principles Wips Trends & data Tools & metrics Countries About

Home > National strategies & policies

National AI policies & strategies

This section provides a live repository of over 1000 AI policy initiatives from 69 countries, territories and the EU. Click on a country/territory, a policy instrument or a group targeted by the policy.

Countries & territories | Policy instruments | Target Groups

Search for a specific dash

Download all AI policies

African Union
Argentina
Armenia
Australia
Austria
Belgium
Brazil
Bulgaria
Canada
Chile
China
Colombia

Costa Rica
Croatia
Cyprus
Czechia
Denmark
Egypt
Estonia
Finland
France
Germany
Greece
Hungary

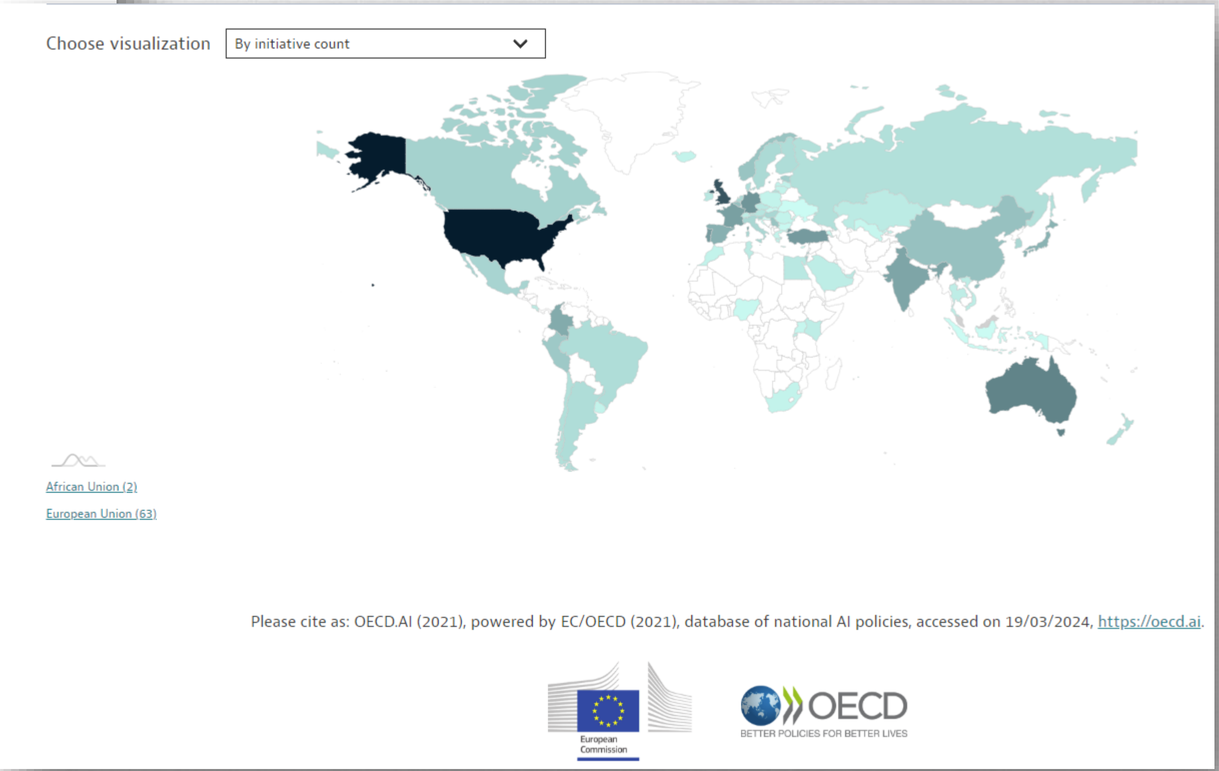
Iceland
India
Indonesia
Ireland
Israel
Italy
Japan
Kazakhstan
Kenya
Korea
Latvia
Lithuania

Luxembourg
Malta
Mauritius
Mexico
Morocco
Netherlands
New Zealand
Nigeria
Norway
Peru
Poland
Portugal

Romania
Rwanda
Saudi Arabia
Serbia
Singapore
Slovakia
Slovenia
South Africa
Spain
Sweden
Switzerland
Thailand

Tunisia
Türkiye
Uganda
Ukraine
United Arab Emirates
United Kingdom
United State
Uruguay
Uzbekistan
Viet Nam
European Un

OVER 1000 AI POLICY INITIATIVES FROM 69 COUNTRIES, TERRITORIES AND THE EU



Choose visualization: By initiative count

African Union (2)

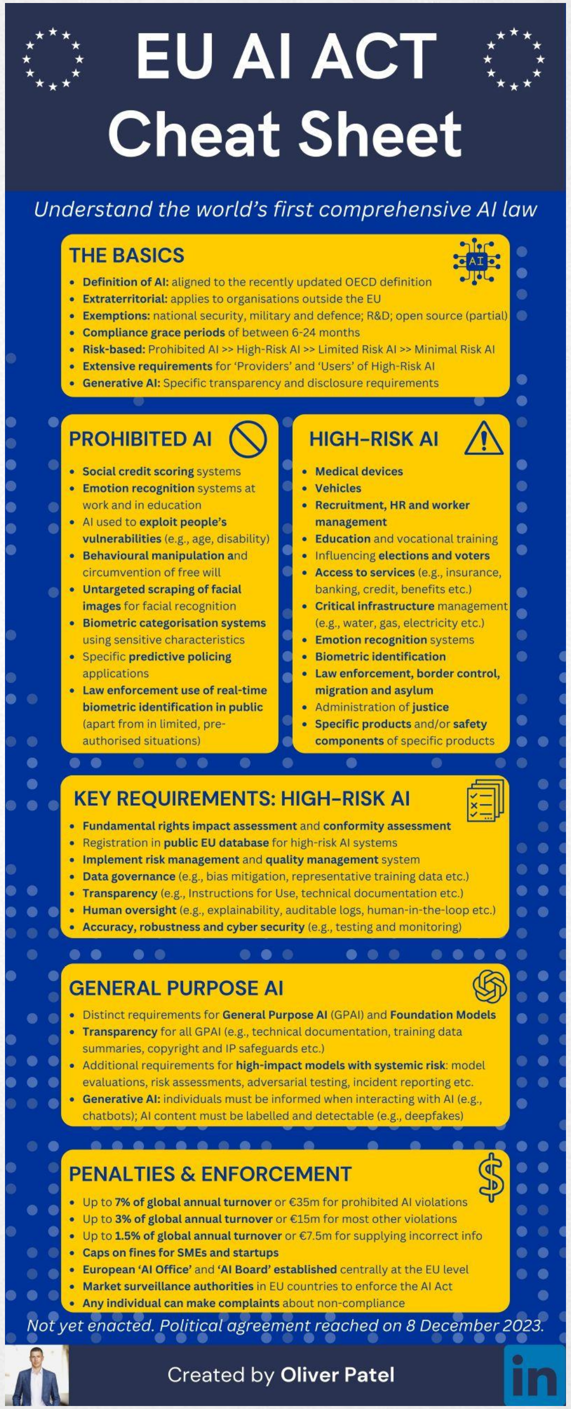
European Union (63)

Please cite as: OECD.AI (2021), powered by EC/OECD (2021), database of national AI policies, accessed on 19/03/2024, https://oecd.ai.

U.S. CURRENTLY HAS 82 INITIATIVES

Additional evolutions in the regulatory environment: Executive orders, councils (NAIAC), Chief AI Officers at Federal Agencies, enforcements, etc.

WORLD’S FIRST COMPREHENSIVE AI LAW



EU AI ACT Cheat Sheet

Understand the world’s first comprehensive AI law

THE BASICS

- **Definition of AI:** aligned to the recently updated OECD definition
- **Extraterritorial:** applies to organisations outside the EU
- **Exemptions:** national security, military and defence; R&D; open source (partial)
- **Compliance grace periods** of between 6-24 months
- **Risk-based:** Prohibited AI >> High-Risk AI >> Limited Risk AI >> Minimal Risk AI
- **Extensive requirements** for ‘Providers’ and ‘Users’ of High-Risk AI
- **Generative AI:** Specific transparency and disclosure requirements

PROHIBITED AI

- **Social credit scoring** systems
- **Emotion recognition** systems at work and in education
- AI used to **exploit people’s vulnerabilities** (e.g., age, disability)
- **Behavioural manipulation a**nd circumvention of free will
- **Untargeted scraping of facial images** for facial recognition
- **Biometric categorisation systems** using sensitive characteristics
- Specific **predictive policing** applications
- **Law enforcement use of real-time biometric identification in public** (apart from in limited, pre-authorised situations)

HIGH-RISK AI

- **Medical devices**
- **Vehicles**
- **Recruitment, HR and worker management**
- **Education** and vocational training
- Influencing **elections and voters**
- **Access to services** (e.g., insurance, banking, credit, benefits etc.)
- **Critical infrastructure** management (e.g., water, gas, electricity etc.)
- **Emotion recognition** systems
- **Biometric identification**
- **Law enforcement, border control, migration and asylum**
- Administration of **justice**
- **Specific products** and/or **safety components** of specific products

KEY REQUIREMENTS: HIGH-RISK AI

- **Fundamental rights impact assessment** and **conformity assessment**
- Registration in **public EU database** for high-risk AI systems
- **Implement risk management** and **quality management** system
- **Data governance** (e.g., bias mitigation, representative training data etc.)
- **Transparency** (e.g., Instructions for Use, technical documentation etc.)
- **Human oversight** (e.g., explainability, auditable logs, human-in-the-loop etc.)
- **Accuracy, robustness and cyber security** (e.g., testing and monitoring)

GENERAL PURPOSE AI

- Distinct requirements for **General Purpose AI** (GPAI) and **Foundation Models**
- **Transparency** for all GPAI (e.g., technical documentation, training data summaries, copyright and IP safeguards etc.)
- Additional requirements for **high-impact models with systemic risk**: model evaluations, risk assessments, adversarial testing, incident reporting etc.
- **Generative AI:** individuals must be informed when interacting with AI (e.g., chatbots); AI content must be labelled and detectable (e.g., deepfakes)

PENALTIES & ENFORCEMENT

- Up to **7% of global annual turnover** or €35m for prohibited AI violations
- Up to **3% of global annual turnover** or €15m for most other violations
- Up to **1.5% of global annual turnover** or €7.5m for supplying incorrect info
- **Caps on fines for SMEs and startups**
- **European ‘AI Office’** and **‘AI Board’ established** centrally at the EU level
- **Market surveillance authorities** in EU countries to enforce the AI Act
- **Any individual can make complaints** about non-compliance

Not yet enacted. Political agreement reached on 8 December 2023.

Created by **Oliver Patel**

EXAMPLES OF ENFORCEMENTS & LEGAL ACTION

CONSUMER DECISION TOOLS

Potential Bias in AI
Consumer Decision Tools
Eyed by FTC, CFPB

DATA SCRAPING

Clearview AI breached
Australians' Privacy

AUTOMATED DECISIONS

Landmark Legal Case
Highlights Corporate
Accountability For
Automated Decisions

CHILDREN'S DATA

FTC Takes Action Against Company
Formerly Known as Weight Watchers
for Illegally Collecting Kids' Sensitive
Health Data

BIOMETRIC INFORMATION

Johnson & Johnson Class
Action Claims Neutrogena
Skin360 Violates BIPA
Laws

INDUSTRY STANDARDS & BEST PRACTICES

AI GOVERNANCE

A system of policies, practices, and processes organizations implement to manage and oversee their use of AI technology and associated risks to ensure the AI aligns with an organization's objectives, is developed and used responsibly and ethically, and complies with applicable legal requirements.

–IAPP Glossary

NIST AI RMF	OECD AI PRINCIPLES	EU AI ACT	MICROSOFT	NYMITY	PWC RESPONSIBLE AI TOOLKIT	OTHER FRAMEWORKS REVIEWED
Valid & Reliable	Robustness	Technical Robustness	Reliability	Robustness	Robustness	EO 13960
Safe	Safety	Safety	Safety	Human Oversight and Promotion of Human Values	Safety	G20 AI Principles
Fair & Bias Managed	Human-centered Values and Fairness	Non-discrimination, Diversity & Fairness, Data Governance	Fairness & Inclusiveness	Diversity, Non-discrimination, and Fairness	Bias & Fairness	Salesforce
Secure & Resilient	Security	Security & Resilience	Security	Security	Security	Jurisdiction Specific Declarations and Regulations
Transparent & Accountable	Transparency and Responsible Disclosure & Accountability	Transparency, Accountability, Human Agency & Oversight	Transparency & Accountability	Transparency, Accountability & Auditability	Data and AI Ethics Policy & Regulation	
Explainable & Interpretable	Explainability		Explainability & Interpretability (as part of Transparency)	Explainability	Interpretability & Explainability	
Privacy-Enhanced	Human values: Respect for Human Rights	Privacy & Data Governance	Privacy	Privacy Data Governance Legality, Necessity and Proportionality	Privacy Governance, Compliance, and Risk Management	

AI GOVERNANCE FRAMEWORKS COMPARISON SUMMARY

Key observation:

Most prominent frameworks include similar core principles:

- Valid & Reliable
- Safe
- Fairness & Bias
- Secure & Resilient
- Transparent & Accountable
- Explainable & Interpretable
- Privacy-Enhanced

GAINS & RISKS

POTENTIAL GAINS

WHEN AI GOES RIGHT

SECURITY

Enhance threat detection and increase system resilience

INNOVATION

Innovate products and services

BUSINESS INTEGRITY

Increase fraud prevention and detection

FINANCIAL

Achieve cost savings, improve demand forecasting or financial projections

OPERATIONAL

Operate more efficiently and increase productivity, improve supply chain and other operation resilience

PEOPLE

Increase skills, retain top talent, enhance recruiting

DECISION-MAKING

Improve decision-making, workforce or scenario planning

QUALITY

Create better customer experiences

REAL WORLD EXAMPLES

ARTIFICIAL INTELLIGENCE IN THE REAL WORLD

PERSONALIZED LEARNING

Rensselaer Polytechnic Institute uses AI for personalized learning in the "Mandarin Project" which creates a lifelike experience to practice the language, in their New York based institution.

ADMINISTRATIVE SUPPORT

The University of Bridgeport suggests using AI to streamline administrative processes to power student record systems, scheduling, etc. AI tools are used to interpret data on recruitment, admissions, and retention efforts.

RESEARCH

Richard Ross, an associate professor at the University of Virginia uses AI in his classroom by having his students conduct research and then compare the results to research done by an AI model.

POTENTIAL CONSEQUENCES & IMPACTS

WHEN AI GOES WRONG

REPUTATION

- Adverse publicity or PR crisis
- Damage to brand reputation
- Loss of trust

LICENSE TO OPERATE

- Revoked licenses
- Sanctions or penalties

LEGAL & REGULATORY

- Enforcements - Fines, penalties, delete AI & data
- Consent decrees or consent agreements
- Private right of action

ENVIRONMENTAL

- Harm to environment
- Wasted resources

FINANCIAL

- Fines
- Financial loss or performance
- Recovery expenses (legal, breach notification, etc.)
- Sunk costs
- Loss of donors

HEALTH & SAFETY

- Harm to individuals, communities, or society

OPERATIONAL

- Apps removed from app stores
- Resource constraints
- Loss of productivity
- Delete algorithms and corresponding data
- Loss of innovation

PEOPLE

- Loss of “customers” (website visitors, students, alumni, survey respondents, etc.), employees, job applicants, top talent
- Negative impact on morale

BUSINESS INTEGRITY/FRAUD

- Complaints tied to inaccurate product claims, untruthful advertising, unfair or deceptive trade practices
- Claims of discrimination or violated individual rights

SALES

- Loss of revenue
- Decline in loyalty or long-term customers, subscribers, or repeat visitors

EXAMPLE

“The other day a colleague asked the ChatGPT website, Who is Sheri Dew? Within seconds, it generated a bio on me that got my birthday, birthplace, and Church membership right. Then it said I was Time Magazine’s 2003 Woman of the Year. I wasn’t. And that I have an MBA from Harvard. I don’t. But it all looked true.

Another colleague in the room said, “I didn’t know you were Time Magazine’s Woman of the Year.” “I wasn’t,” I said for the second time. Trust me, I am going to get introduced somewhere as Time’s Woman of the Year. Artificial intelligence has remarkable capabilities, but it can also produce deepfakes instantly…
Artificial Intelligence will only make it more difficult to discern what is true and what is not. “

–SHERI L. DEW

BYU Women’s Conference, May 5, 2023

REAL WORLD EXAMPLES

ARTIFICIAL INTELLIGENCE IN THE REAL WORLD

MISINFORMATION

Google's AI Overviews feature has delivered numerous incorrect answers, such as:

-Batman is a cop

-A dog has played in the NBA, NFL, and NHL

-Add glue to your pizza in order to get the cheese to stay on.

HIRING

Amazon stopped using its AI hiring algorithm after finding it favored applicants based on the use of certain words more commonly found in men's resumes.

FIRED & FINED

One lawyer was fired and another fined after using ChatGPT to save time. In both cases, the AI chatbot made up several fake lawsuit citations when asked to compose a brief that were not discovered until after filing.

RECENT HEADLINES

EXAMPLES OF REPUTATIONAL DAMAGE, HARM TO INDIVIDUALS, AND LOSS OF TRUST

THEY THOUGHT LOVED ONES WERE CALLING FOR HELP. IT WAS AN AI SCAM

-The Washington Post
May 5, 2023

TEEN BOYS ACCUSED OF CREATING AI DEEPFAKE NUDES OF FEMALE CLASSMATES

-The Independent
Nov 2, 2023

MICROSOFT ACCUSED OF DAMAGING GUARDIAN'S REPUTATION WITH AI-GENERATED POLL

-The Guardian
Oct 31, 2023

IS YOUR HEALTH INSURER USING AI TO DENY YOU SERVICES? LAWSUIT SAYS ERRORS HARMED ELDERS

-USA Today
Nov 19, 2023

DON'T DATE ROBOTS — THEIR PRIVACY POLICIES ARE TERRIBLE

-The Verge
Feb 15, 2024

SPORTS ILLUSTRATED IS THE LATEST MEDIA COMPANY DAMAGED BY AN AI EXPERIMENT GONE WRONG

-AP News
Nov 28, 2023

MAN TRICKS GM DEALER'S AI CHATBOT INTO SELLING 2024 CHEVY TAHOE FOR $1

-Breitbart
Dec 22, 2023

FACIAL RECOGNITION LED TO WRONGFUL ARRESTS. SO DETROIT IS MAKING CHANGES

-The New York Times
June 29, 2024

When AI Systems Fail: Introducing the AI Incident Database - Partnership on AI

Welcome to the Artificial Intelligence Incident Database

WHAT ARE WE DOING ABOUT IT AT BYU?

—

CES PRIVACY CENTER

CORE STRATEGIC COMPONENTS

BYU AI Committee

Defined Mission & Role

Cross-Functional Collaboration

Standards & Resources Governing & Facilitating Employee Use

Guiding Principles

Training and Awareness

ROLE OF BYU AI COMMITTEE

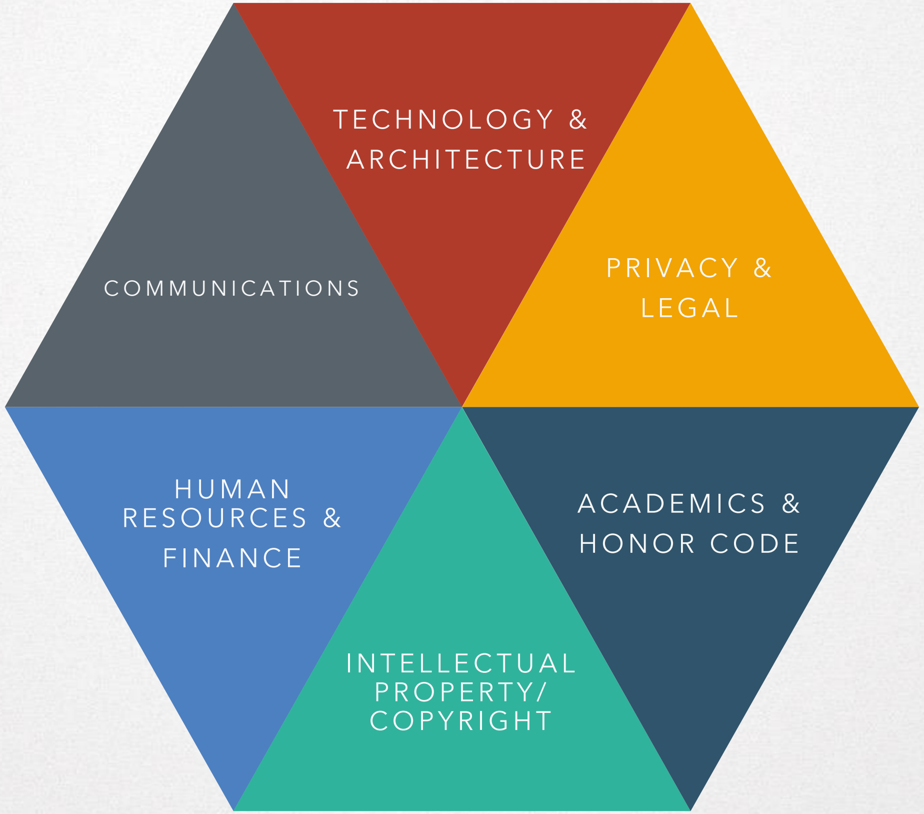
The AI Committee helps guide the university in understanding and navigating the rapidly evolving trends in AI technology, including Generative AI, to promote safe and strategic adoption. It is tasked with analyzing AI's impact, developing policy recommendations, bolstering the university's grasp of AI's potential, and ensuring practices are informed, consistent and spiritually strengthening.

Additionally, the committee will establish and maintain a central repository for AI guidelines, resources, and tools to support uniform application and promote safety and best practices across the university.

BYU

BRIGHAM YOUNG UNIVERSITY

CROSS-FUNCTIONAL COLLABORATION



TECHNOLOGY & ARCHITECTURE

COMMUNICATIONS

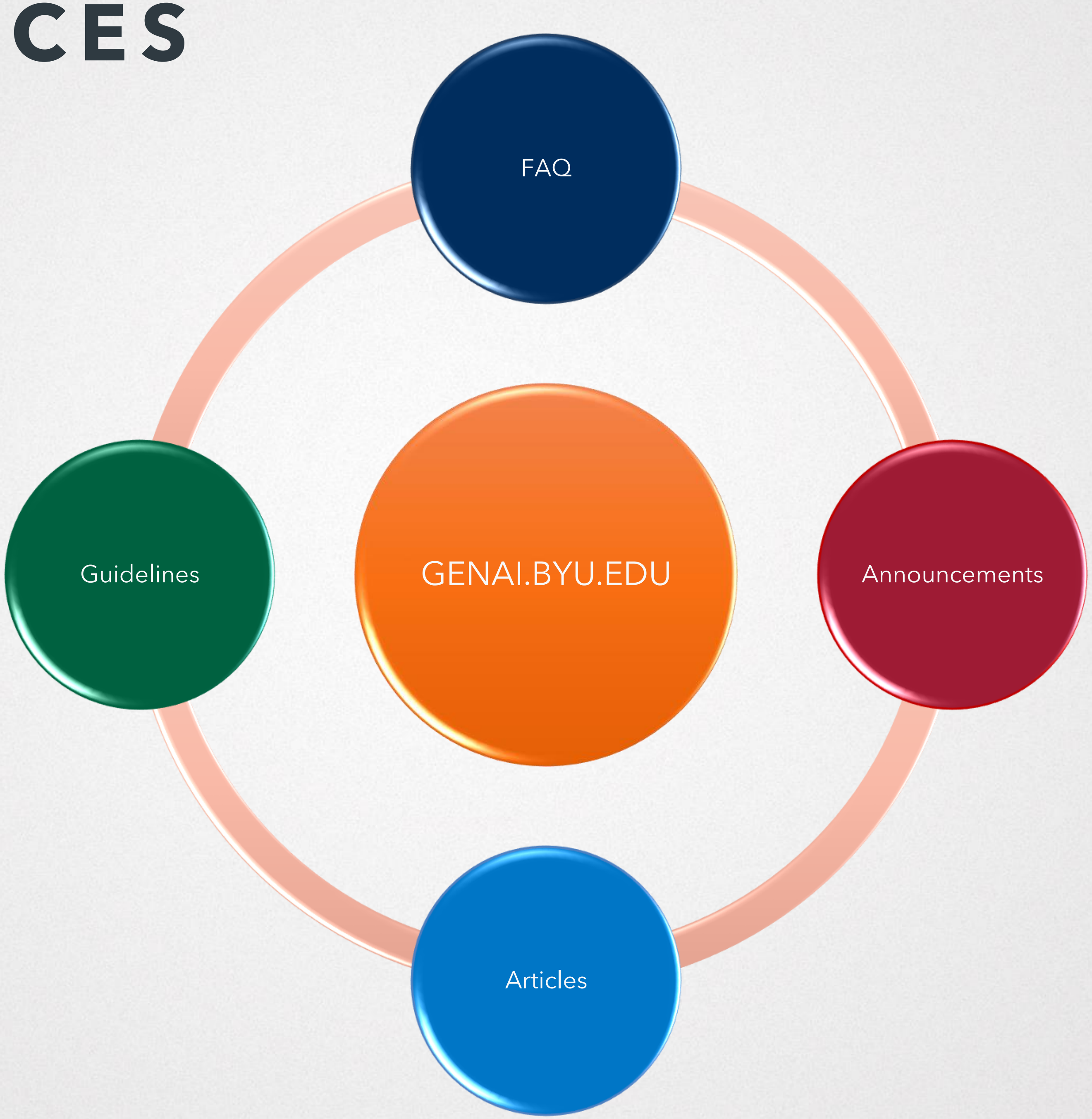
PRIVACY & LEGAL

HUMAN RESOURCES & FINANCE

ACADEMICS & HONOR CODE

INTELLECTUAL PROPERTY/ COPYRIGHT

RESOURCES



GUIDING PRINCIPLES

Integrity

Transparency

Data Protection

Accountability

WHAT DO YOU NEED TO DO?

CES PRIVACY CENTER

START BY APPLYING THE GUIDING AI PRINCIPLES

INTEGRITY

Includes integrity in our **data**, our **processes**, and **as individuals**. It involves maintaining the quality and accuracy of data inputs and outputs when using AI systems. This principle requires us to develop and apply standardized processes and rules as we test and operate AI systems. It involves measures to mitigate risk, prevent data hallucinations, and to avoid biases and inaccuracies. Integrity requires a steadfast commitment to acting in ethical and inclusive ways to promote fairness and compliance. It also includes using AI technologies to promote our educational and spiritual values and objectives to enhance lives in positive and uplifting ways.

KEY TAKEAWAYS:

- Promote educational and spiritual values and objectives.
- Prioritize data quality and accuracy.
- Develop and apply standardized processes.

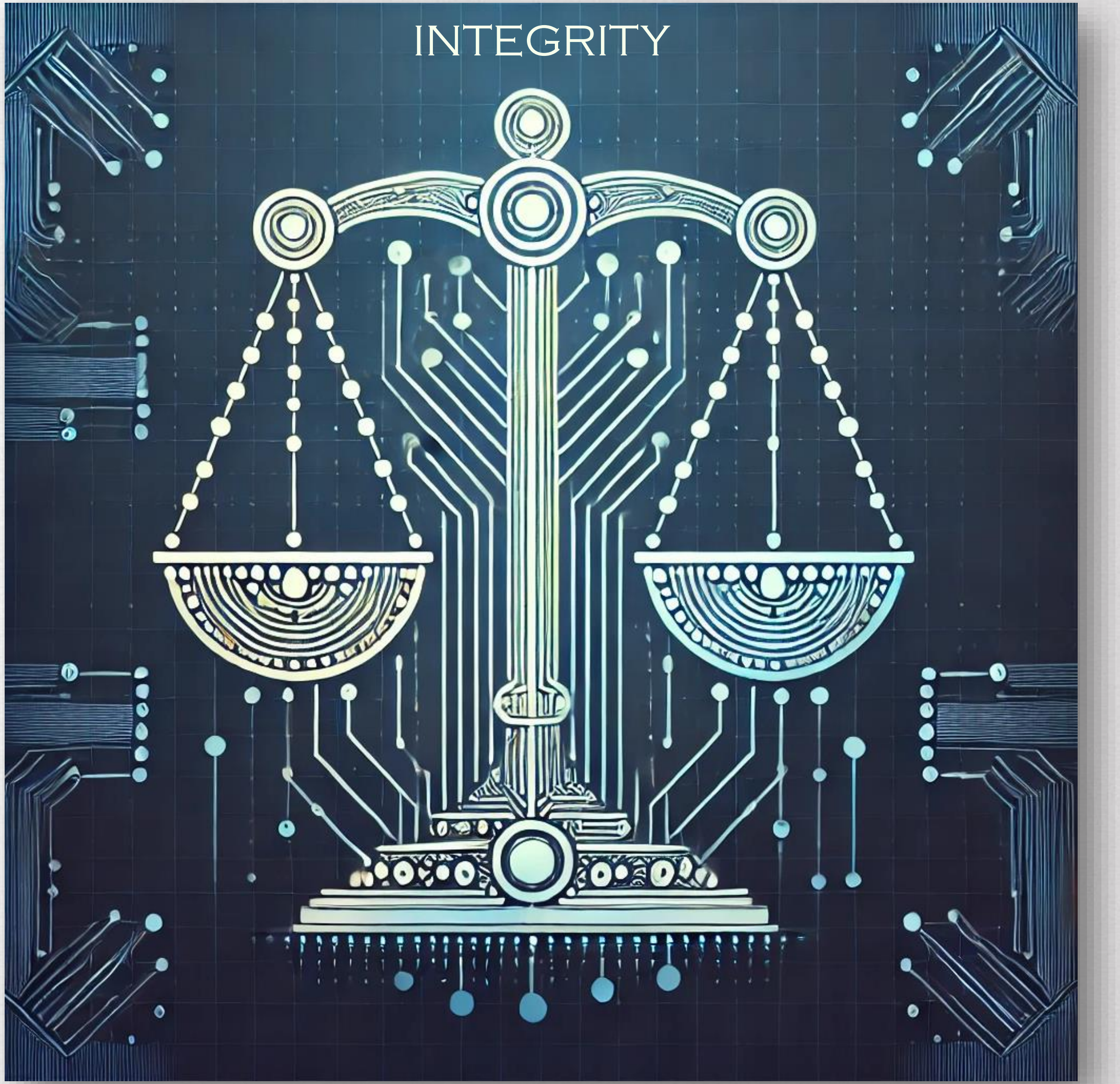


Image partially generated by ChatGPT

DATA PROTECTION

Emphasizes both **security** and **privacy**. It encompasses the safeguarding of personal data against unauthorized access, use, and loss. This involves implementing robust security measures, adhering to data privacy laws (such as GDPR and FERPA), and ensuring that personal data is collected, used, and stored only for specified, legitimate purposes. Processing data in line with **data governance** standards and **retention** policies helps us handle it in a way that respects individual privacy rights and mitigates risk. Data Protection also includes prioritizing safety, avoiding harm to individuals, and using AI systems that are resilient and reliable.

KEY TAKEAWAYS:

- Embed security, privacy, data governance, data retention.
- Prevent and mitigate risk.

Image partially generated by ChatGPT

TRANSPARENCY

Is about making the AI systems' functionalities, data usage, and decision-making processes **open** and **understandable** to users and stakeholders. It entails providing clear, intelligible information about how AI systems operate, when they are in use, the data they use, and the logic behind their decisions. This principle supports the right to explanation, **enabling individuals to understand** and, if necessary, challenge AI-driven decisions.

KEY TAKEAWAYS:

- Be honest and open.
- Provide clear information about when AI is in use.

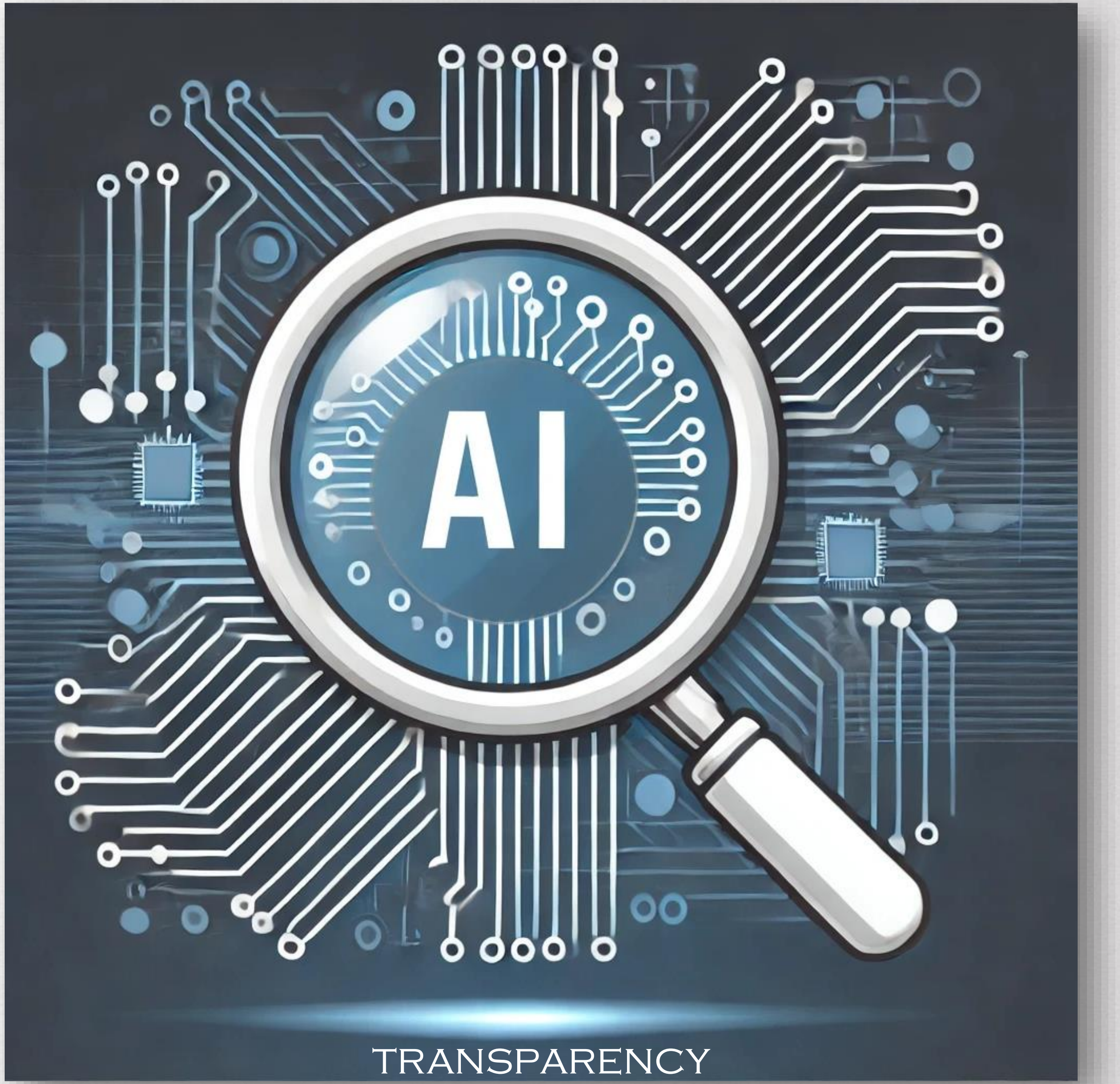


Image partially generated by ChatGPT

ACCOUNTABILITY

Entails our commitment to be **answerable for the outcomes** of the AI systems we use. This includes establishing clear **governance** structures that define roles and responsibilities within the institution for AI decision-making and ensuring that AI systems are used in **compliance** with **ethical standards** and legal requirements. Accountability mechanisms may include **policies, procedures**, internal audits, trainings, and the establishment of a committee to oversee AI practices.

KEY TAKEAWAYS:

- Adhere to policies, standards, and procedures.
- Be responsible for how we use AI.

Image partially generated by ChatGPT

AI GUIDING PRINCIPLES

Whenever you use AI, **always remember to apply these 4 guiding AI principles**. Together, these principles form a foundation for the responsible governance of AI, promoting the use of AI technologies in ways that are ethical, legal, respectful of privacy and security, and aligned with our institutional values and goals, all while we enable innovation, gain efficiency, and harness even greater untapped potential.

1. Integrity

- Promote educational and spiritual values and objectives.
- Prioritize data quality and accuracy.
- Develop and apply standardized processes.

2. Data Protection

- Embed security, privacy, data governance, and data retention.
- Prevent and mitigate risk.

3. Transparency

- Be honest and open.
- Provide clear information about when AI is in use.

4. Accountability

- Adhere to policies, standards, and procedures.
- Be responsible for how we use AI.

AI GUIDING PRINCIPLES: SELF-CHECK

For additional guidance, below is a quick checklist to assist you in adhering to the four AI Guiding Principles.

1. Integrity

- □ Use AI to promote educational and spiritual values and objectives, enhancing lives in positive and uplifting ways.
- □ Increase awareness of the strengths and limitations of the AI tool to set realistic expectations on how you can leverage it. Use it to supplement, not replace human work.
- □ Review data outputs to verify accuracy, relevancy, and data quality.
    - □ Always cross-check AI-generated results to make sure they are accurate and appropriate.
- □ Periodically evaluate the AI tool's performance using diverse and updated datasets to identify and address potential biases or inaccuracies.

2. Data Protection

Security

- □ Keep your device and application versions updated to prevent vulnerabilities.
- □ Ensure that if you use mobile app versions, they are kept up to date as well.
- □ Use Multi-Factor Authentication (MFA) to enhance account security.
- □ Only use well-vetted and approved AI tools.

Privacy

- □ Review the privacy notice and data handling/sharing practices of any AI tool before usage.
- □ Only use well-vetted and approved AI tools.
- □ Do not share data with any AI tool without a business contract to ensure data privacy.
- □ Do not expose sensitive data or violate privacy regulations.
    - □ Do not enter Nonpublic Institutional Data into any AI tool. This type of data includes personally identifiable employee data, FERPA-covered student data, HIPAA-covered patient data, and may include research that is not yet publicly available. Refer to BYU's Data Use, Privacy, and Security Policy for more information.
- □ Maintain the strictest default privacy settings within the AI Tool.
    - □ Do not allow your data or conversations to be used or stored to improve or train models when possible. If this is a requirement for usage, question if usage is strictly necessary.
- □ Stay informed about data privacy best practices, by participating in institutional trainings, asking questions, and learning together about this constantly evolving space.

Data Retention

- □ If feasible, delete chats or threads within an AI tool once they have fulfilled their intended purpose.
- □ Follow applicable data retention policies and standards for all data outputs.

3. Transparency

- □ Cite the use of AI tools when applicable.
- □ Maintain documentations of data handling, chatbot training protocols, and decision-making processes when appropriate and when used to make decisions that influence individuals and the institution.
    - □ Provide individuals insight into AI-enabled decisions when requested.

4. Accountability

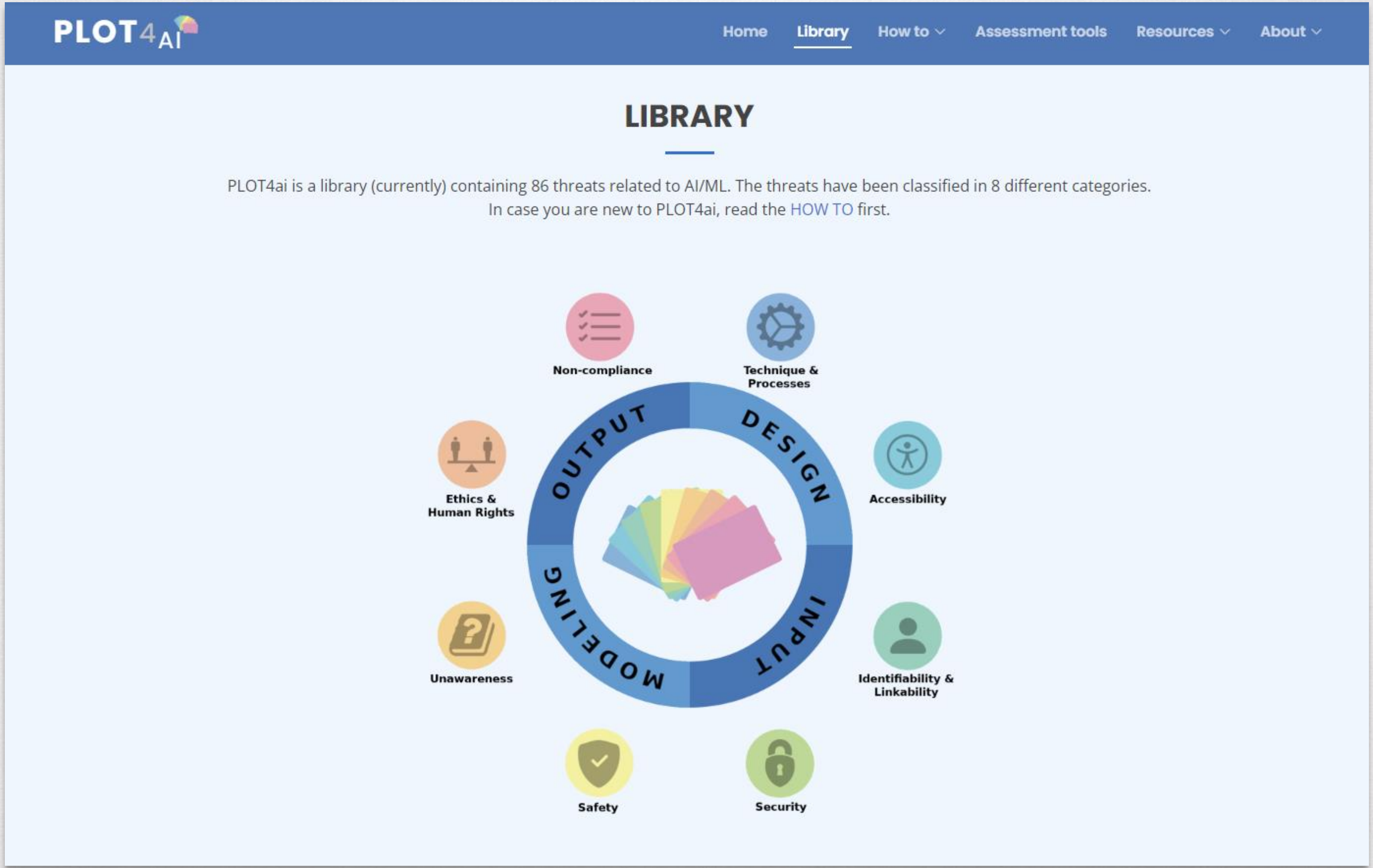
- □ Take time to understand how the AI tool or software works and know its limitations.
- □ Use AI tools in appropriate and ethical ways, aligned to our institution's values and objectives.
    - □ Only input data and prompts that would not cause harm to an individual or the institution.
- □ Restrict access to customized GPTs or similar chatbots and data outputs based on user roles and responsibilities.
- □ When in doubt of what is appropriate use; ask, do not assume.
    - □ Ask your department CSR or OIT representative for clarification on AI usage or reach out to the CES Privacy Center for Privacy or AI Governance related concerns.
- □ Adhere to the existing policies, standards, and procedures put in place by the institution.
    - □ A few key ones are Data, Use, Privacy, and Security; Privacy Notice; Academic Integrity Policy; and CES Honor Code
- □ Stay informed. Participate in trainings to increase awareness and understanding of both potential risks and gains tied to the usage of various AI tools.

AI GUIDING PRINCIPLES & SELF-CHECK

ADDITIONAL RESOURCES

CES PRIVACY CENTER

EXPLORING RISKS IN AI:

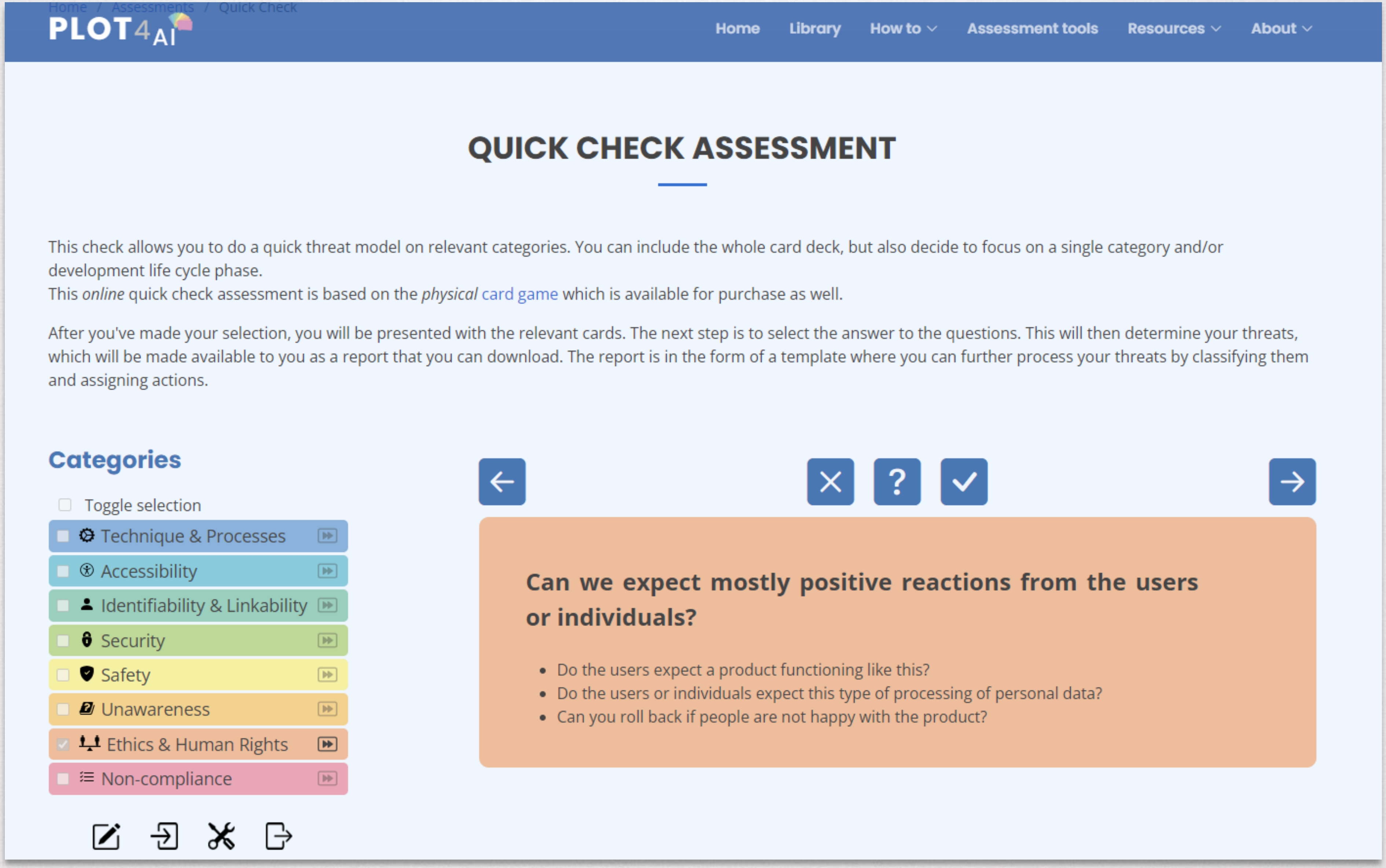


LIBRARY

PLOT4ai is a library (currently) containing 86 threats related to AI/ML. The threats have been classified in 8 different categories.
In case you are new to PLOT4ai, read the HOW TO first.

https://plot4.ai/library

EXAMPLE ASSESSMENT:



QUICK CHECK ASSESSMENT

This check allows you to do a quick threat model on relevant categories. You can include the whole card deck, but also decide to focus on a single category and/or development life cycle phase.
This *online* quick check assessment is based on the *physical* card game which is available for purchase as well.

After you've made your selection, you will be presented with the relevant cards. The next step is to select the answer to the questions. This will then determine your threats, which will be made available to you as a report that you can download. The report is in the form of a template where you can further process your threats by classifying them and assigning actions.

Categories

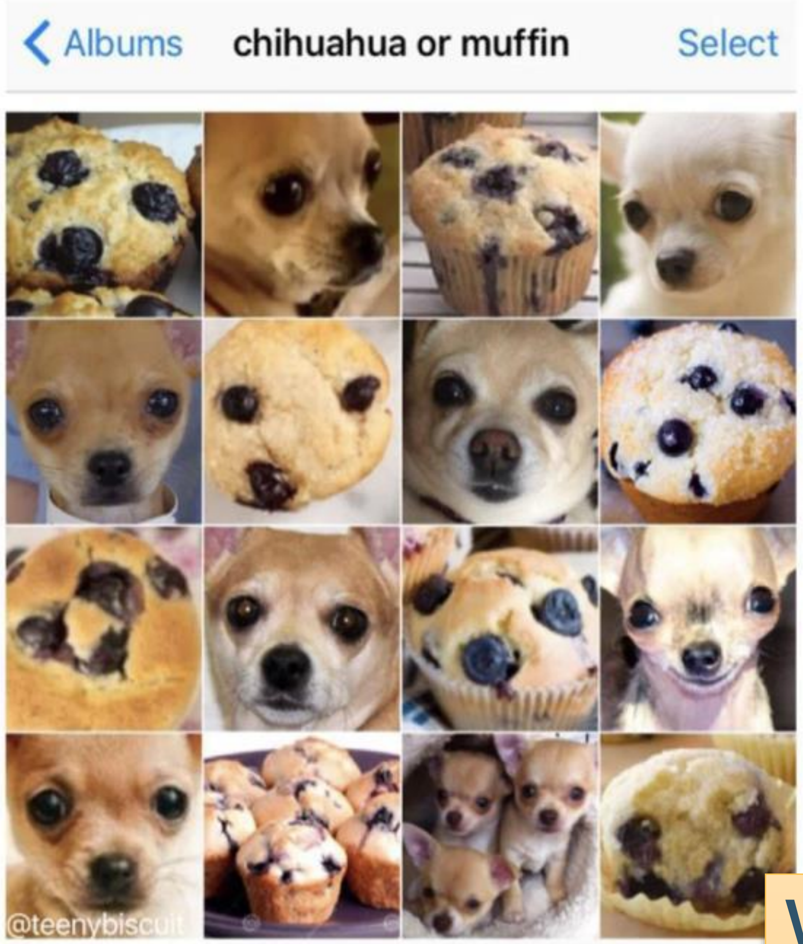
- □ Toggle selection
- □ Technique & Processes
- □ Accessibility
- □ Identifiability & Linkability
- □ Security
- □ Safety
- □ Unawareness
- ☑ Ethics & Human Rights
- □ Non-compliance

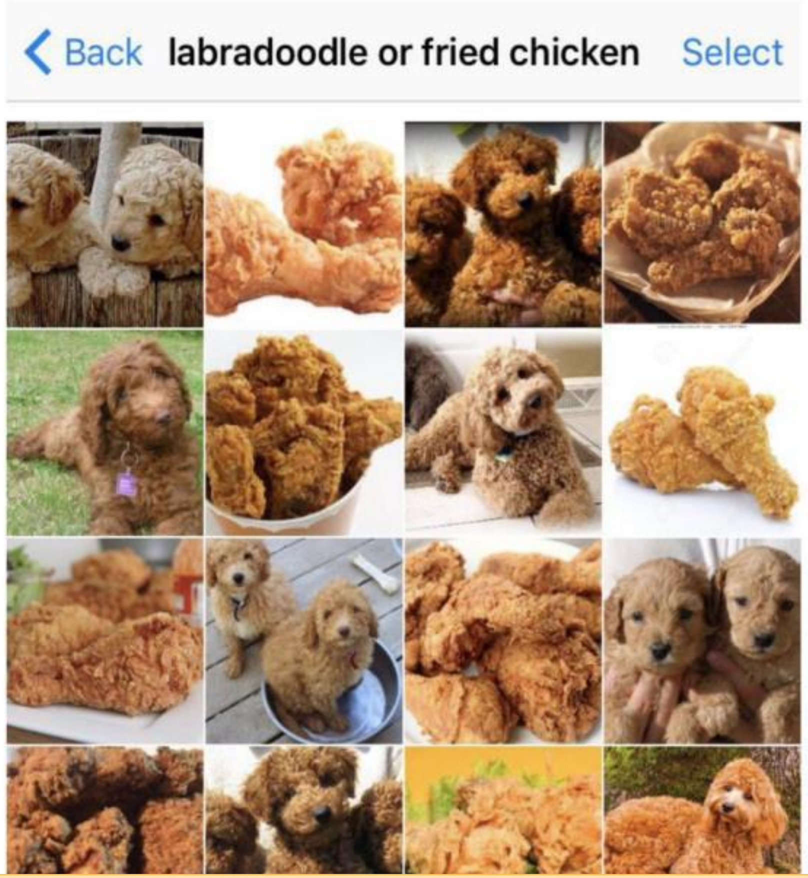
Can we expect mostly positive reactions from the users or individuals?

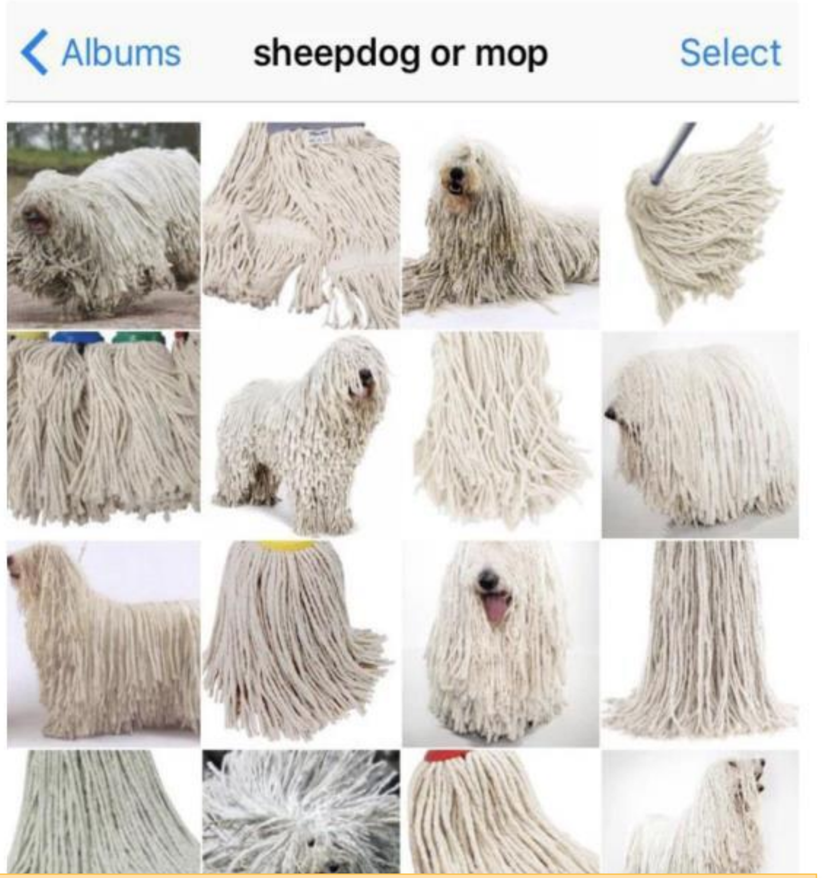
- Do the users expect a product functioning like this?
- Do the users or individuals expect this type of processing of personal data?
- Can you roll back if people are not happy with the product?

https://plot4.ai/library

EXPLORING ACCURACY IN AI:







Images are from an AI class held at the IAPP's Privacy. Security. Risk Conference taught by:

Dr. Sara Jordan Senior Researcher, *Artificial Intelligence and Ethics at the Future of Privacy Forum*

Ilana Golbin Blumenfeld, *Global Responsible AI Lead, PwC*

What do we do when AI is inaccurate?

"Employees should be regularly reminded that: generative AI outputs can be incorrect, out-of-date, biased, or misleading. Individuals are responsible for the content they create, regardless of the assistance of generative AI tools, and employees are encouraged to independently verify the accuracy of any outputs. Verification is particularly important when employees use AI in situations that require legal certification of accuracy, e.g. financial reports, court filings, and due diligence documents."

- Amber Ezzell, *Future of Privacy Forum, Generative AI for Organizational Use: Internal Policy Checklist, July 2023*

EXPLORING BIAS IN AI:

Exercise 1: Search “Doctor” vs “Nurse” in Google images.

Exercise 2: Search “Beautiful” in Google images.

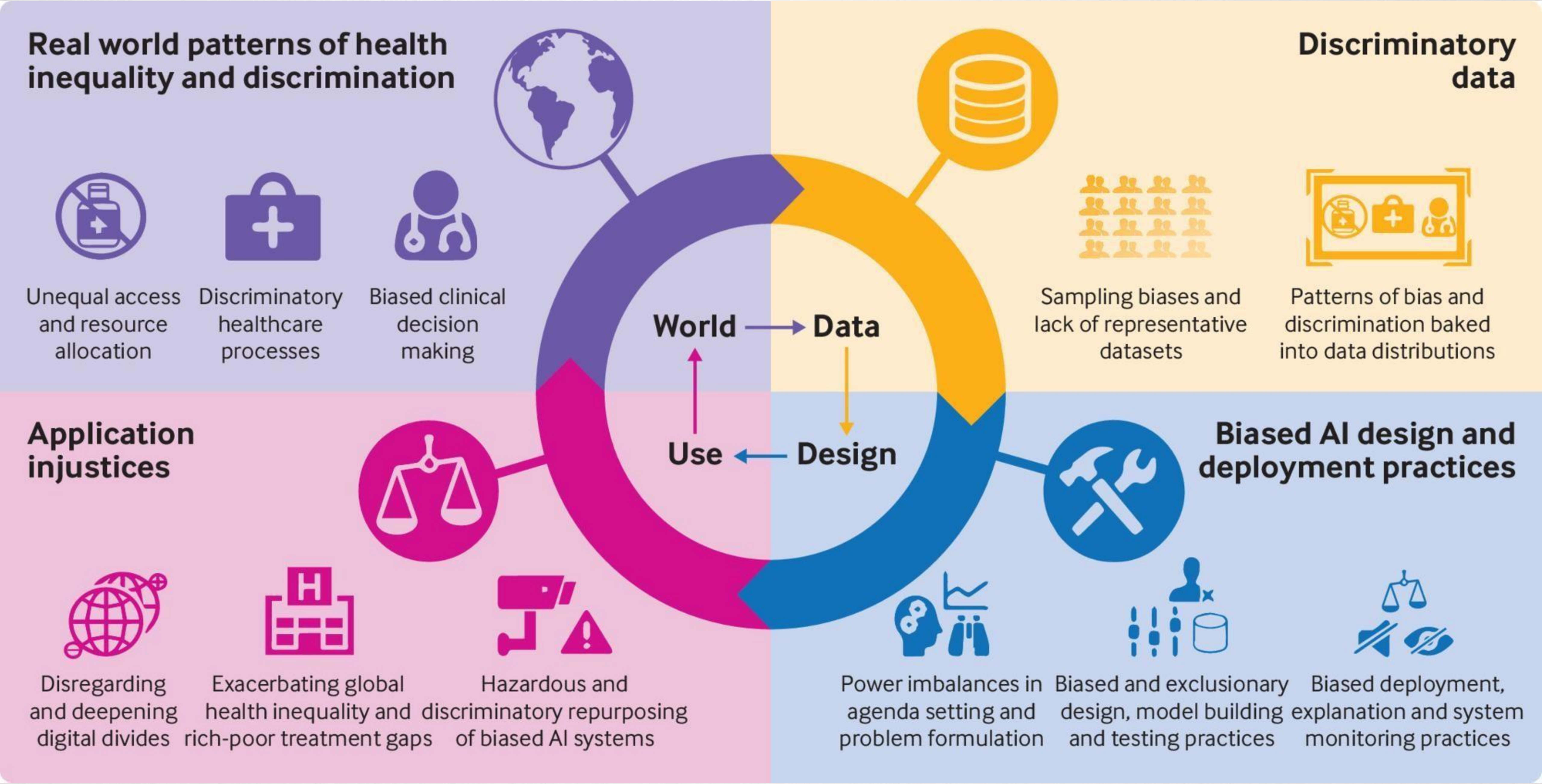
How does bias occur in AI?

“AI bias is caused by bias in data sets, people designing AI models and those interpreting its results.

People write the algorithms; people choose the data used by algorithms and people decide how to apply the results of the algorithms. Without diverse teams and rigorous testing, it can be too easy for people to let subtle, unconscious biases enter, which AI then automates and perpetuates.”

- PWC Responsible AI Toolkit

EXAMPLE OF BIAS IN AI:



Real world patterns of health inequality and discrimination

Unequal access and resource allocation

Discriminatory healthcare processes

Biased clinical decision making

Discriminatory data

Sampling biases and lack of representative datasets

Patterns of bias and discrimination baked into data distributions

Application injustices

Disregarding and deepening digital divides

Exacerbating global health inequality and rich-poor treatment gaps

Hazardous and discriminatory repurposing of biased AI systems

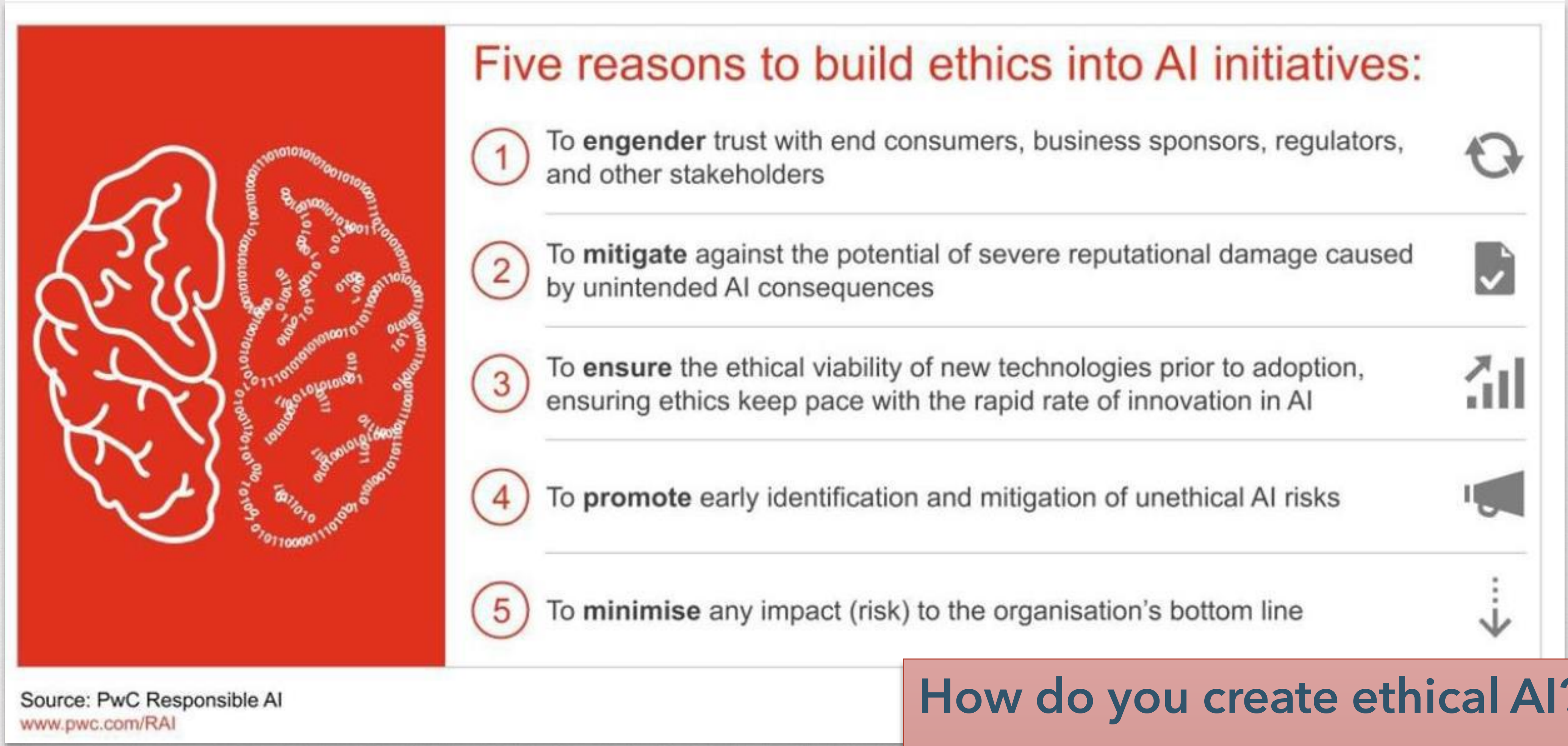
Biased AI design and deployment practices

Power imbalances in agenda setting and problem formulation

Biased and exclusionary design, model building and testing practices

Biased deployment, explanation and system monitoring practices

EXPLORING ETHICS IN AI:



Five reasons to build ethics into AI initiatives:

1. To **engender** trust with end consumers, business sponsors, regulators, and other stakeholders
2. To **mitigate** against the potential of severe reputational damage caused by unintended AI consequences
3. To **ensure** the ethical viability of new technologies prior to adoption, ensuring ethics keep pace with the rapid rate of innovation in AI
4. To **promote** early identification and mitigation of unethical AI risks
5. To **minimise** any impact (risk) to the organisation’s bottom line

Source: PwC Responsible AI
www.pwc.com/RAI

How do you create ethical AI?

“In order to create end-to-end ethics-by-design, mature AI ethics practices combine ethical AI product development and engineering with privacy, legal, user research, design, and accessibility partners to create a holistic approach to the development, marketing, sale, and implementation of AI.”

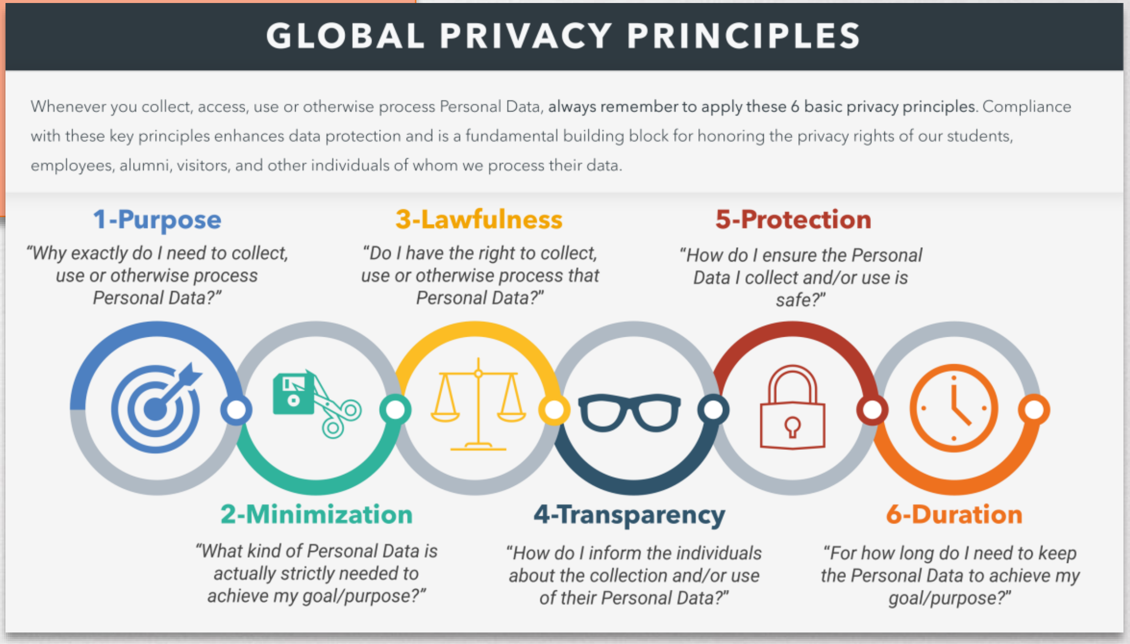
- Kathy Baxter, *Principal Architect, Salesforce Ethical AI Practice*

EXPLORING PRIVACY & SECURITY IN AI:

“One of the principles of responsible AI regularly mentioned refers explicitly to “**privacy**.” This is reminiscent of the obligation to **apply general privacy principles**, which are the backbone of privacy and data protection globally, to AI/ML systems which process personal data. This includes ensuring collection limitation, data quality, purpose specification, use limitation, accountability and individual participation.

Principles of trustworthy AI like transparency and explainability, fairness and non-discrimination, human oversight, robustness and **security** of data processing can regularly be related to specific individual rights and provisions of corresponding privacy laws.”

-Katharina Koerner
https://iapp.org/news/a/privacy-and-responsible-ai/



GLOBAL PRIVACY PRINCIPLES

Whenever you collect, access, use or otherwise process Personal Data, **always remember to apply these 6 basic privacy principles**. Compliance with these key principles enhances data protection and is a fundamental building block for honoring the privacy rights of our students, employees, alumni, visitors, and other individuals of whom we process their data.

1-Purpose
“Why exactly do I need to collect, use or otherwise process Personal Data?”

2-Minimization
“What kind of Personal Data is actually strictly needed to achieve my goal/purpose?”

3-Lawfulness
“Do I have the right to collect, use or otherwise process that Personal Data?”

4-Transparency
“How do I inform the individuals about the collection and/or use of their Personal Data?”

5-Protection
“How do I ensure the Personal Data I collect and/or use is safe?”

6-Duration
“For how long do I need to keep the Personal Data to achieve my goal/purpose?”

KEY AI DEFINTIONS

ARTIFICIAL INTELLIGENCE

A broad term used to describe an engineered system where machines learn from experience, adjusting to new inputs, and potentially performing tasks previously done by humans. More specifically, it is a field of computer science dedicated to simulating intelligent behavior in computers. It may include automated decision-making.

AUTOMATED DECISION-MAKING

The process of making a decision by technological means without human involvement

MACHINE LEARNING

A subfield of AI involving algorithms that enable computer systems to iteratively learn from and then make decisions, inferences or predictions based on data. These algorithms build a model from training data to perform a specific task on new data without being explicitly programmed to do so.

ML implements various algorithms that learn and improve by experience in a problem-solving process that includes data cleansing, feature selection, training, testing and validation. Common examples: fraud detection, recommender systems, customer inquiries, natural language processing, health care, or transport and logistics.

DEEP LEARNING

A subfield of AI and machine learning that uses artificial neural networks. Deep learning is especially useful in fields where raw data needs to be processed, like image recognition, natural language processing and speech recognition.

ALGORITHM

A computational procedure or set of instructions and rules designed to perform a specific task, solve a particular problem, or produce a machine learning or AI model.

AI GOVERNANCE

A system of policies, practices and processes organizations implement to manage and oversee their use of AI technology and associated risks to ensure the AI aligns with an organization's objectives, is developed and used responsibly and ethically, and complies with applicable legal requirements.

GENERATIVE AI

A field of AI that uses machine learning models trained on large data sets to create new content, such as written text, code, images, music, simulations and videos. These models are capable of generating novel outputs based on input data or user prompts.

ADDITIONAL TERMS AND INFO: https://iapp.org/resources/glossary/

KEY AI DEFINTIONS

SUPERVISED LEARNING

A subset of machine learning where the model is trained on input data with known desired outputs. These two groups of data are sometimes called predictors and targets, or independent and dependent variables, respectively.

This type of learning is useful for training an AI to group data into specific categories or making predictions by understanding the relationship between two variables.

UNSUPERVISED LEARNING

A subset of machine learning where the model is trained by looking for patterns in an unclassified data set with minimal human supervision. The AI is provided with preexisting data sets and then analyzes those data sets for patterns.

This type of learning is useful for training an AI for techniques such as clustering data (outlier detection, etc.) and dimensionality reduction (feature learning, principal component analysis, etc.).

REINFORCEMENT LEARNING

A machine learning method that trains a model to optimize its actions within a given environment to achieve a specific goal, guided by feedback mechanisms of rewards and penalties.

This training is often conducted through trial-and-error interactions or simulated experiences that do not require external data. For example, an algorithm can be trained to earn a high score in a video game by having its efforts evaluated and rated according to success toward the goal.

INPUT DATA

Data provided to or directly acquired by a learning algorithm or model for the purpose of producing an output. It forms the basis upon which the machine learning model will learn, make predictions and/or carry out tasks.

SYNTHETIC DATA

Data generated by a system or model that can mimic and resemble the structure and statistical properties of real data. It is often used for testing or training machine learning models, particularly in cases where real-world data is limited, unavailable or too sensitive to use.

TESTING DATA

A subset of the data set used to provide an unbiased evaluation of a final model. It is used to test the performance of the machine learning model with new data at the very end of the model development process.

ADDITIONAL TERMS AND INFO: https://iapp.org/resources/glossary/

THE SPECTRUM OF ARTIFICIAL INTELLIGENCE

Produced by
FUTURE OF PRIVACY FORUM
FPF.ORG

Artificial Intelligence (AI) is the computerized ability to perform tasks commonly associated with human intelligence, including reasoning, discovering patterns and meaning, generalizing, applying knowledge across spheres of application, and learning from experience. The growth of AI-based systems in recent years has garnered much attention, particularly in the sphere of Machine Learning. A subset of AI, Machine Learning (ML) systems "learn" from the success or accuracy of their outputs, and can change their processing over time, with minimal human intervention. But there are non-ML types of AI that, alone or in combination, lie behind the real-world applications in common use. General AI — a human-level computational system — does not yet exist. But Narrow AI exists in many fields and applications where computerized systems greatly enhance human output or outperform humans at defined tasks. This chart explains the main types of AI, their relationships to each other, and provides specific examples of how they are currently appear in our day-to-day lives. It also demonstrates how AI exists within the timeline of human knowledge and development.

AI USE CASES AND CONTEXTS

FINANCE
TAX COMPLIANCE

A software platform that distills tax laws into a program, creates a personalized decision system, and enables individuals to quickly and accurately file their taxes.

Value of AI: Tax compliance requires complete accuracy. This efficient, interactive system that provides precise and logically connected results allows taxpayers to understand, confirm, and have confidence in the outcome. KE provides transparent and clear explanations.

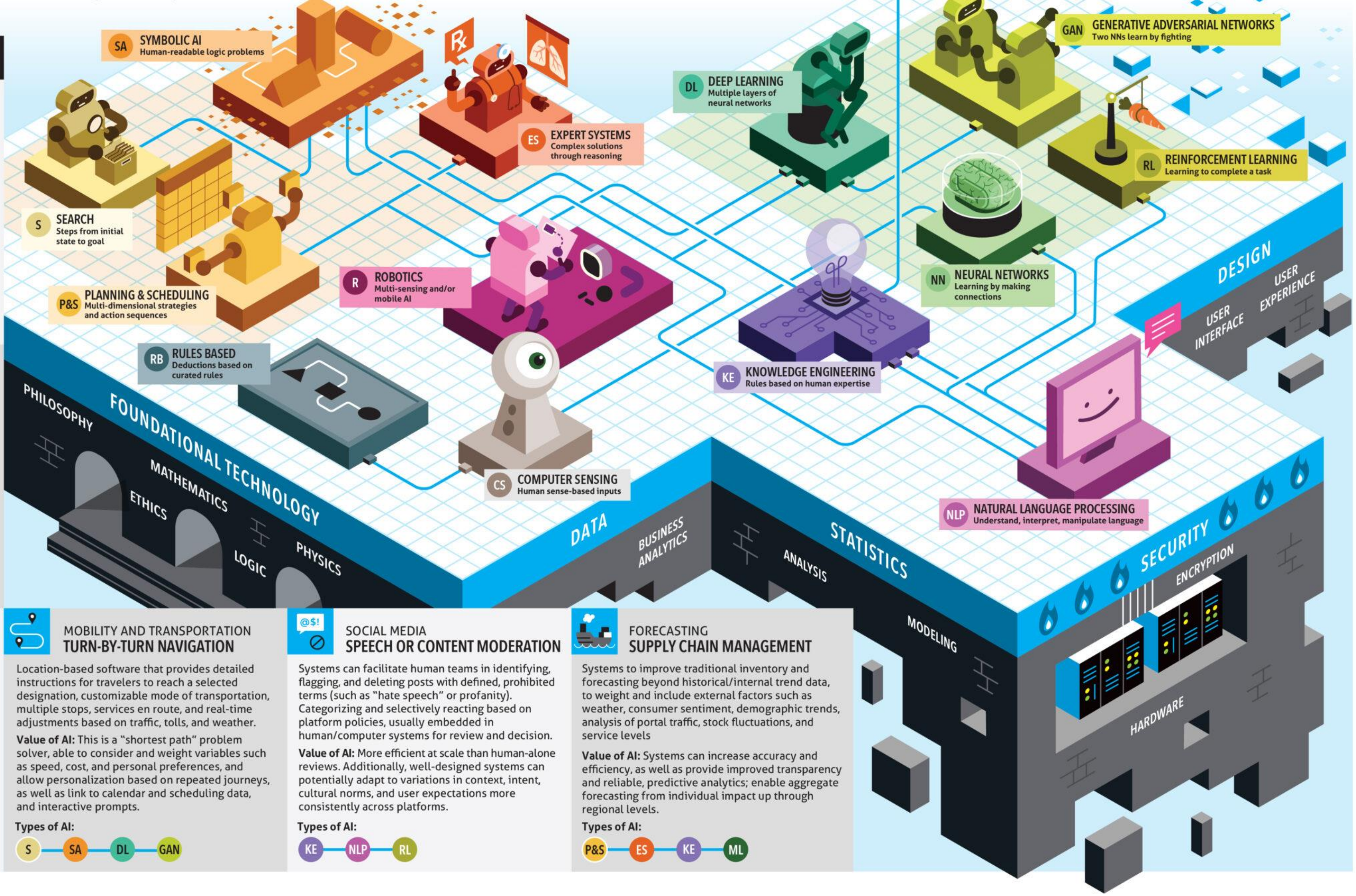
Types of AI: KE — NN — NLP

HEALTHCARE
AMBIENT CHARTING

The use of background voice-to-text processing during a patient/medical provider exchange to record those interactions into the patient's chart, along with extracting tasks, symptoms, and recommendations for further action as required.

Value of AI: Medical providers spend significant time documenting, with uneven outputs, as well as difficulty in correlating between providers. Ambient systems encode conversations, target key phrases, and present a summary for provider edit/acceptance.

Types of AI: SA — DL — NLP


SA SYMBOLIC AI Human-readable logic problems
ES EXPERT SYSTEMS Complex solutions through reasoning
S SEARCH Steps from initial state to goal
P&S PLANNING & SCHEDULING Multi-dimensional strategies and action sequences
R ROBOTICS Multi-sensing and/or mobile AI
RB RULES BASED Deductions based on curated rules
CS COMPUTER SENSING Human sense-based inputs
ML MACHINE LEARNING Algorithms improve through experience
DL DEEP LEARNING Multiple layers of neural networks
GAN GENERATIVE ADVERSARIAL NETWORKS Two NNs learn by fighting
RL REINFORCEMENT LEARNING Learning to complete a task
NN NEURAL NETWORKS Learning by making connections
KE KNOWLEDGE ENGINEERING Rules based on human expertise
NLP NATURAL LANGUAGE PROCESSING Understand, interpret, manipulate language
FOUNDATIONAL TECHNOLOGY
PHILOSOPHY
MATHEMATICS
ETHICS
LOGIC
PHYSICS
DATA
BUSINESS ANALYTICS
STATISTICS
ANALYSIS
MODELING
SECURITY
ENCRYPTION
HARDWARE
DESIGN
USER INTERFACE
USER EXPERIENCE

TRACKING
WORKPLACE MONITORING

Embedded systems can monitor physical and digital traffic, data usage, device management, and some employee behaviors for efficiency and security management of time, assets, and resources.

Value of AI: Monitoring enables necessary enforcement of data security policies and protocols. Also, systems can monitor and manage time reporting and project management tools, as well as ensuring appropriate supervision, training, and support, including for remote workers

Types of AI: RB — CS — NN

MOBILITY AND TRANSPORTATION
TURN-BY-TURN NAVIGATION

Location-based software that provides detailed instructions for travelers to reach a selected designation, customizable mode of transportation, multiple stops, services en route, and real-time adjustments based on traffic, tolls, and weather.

Value of AI: This is a "shortest path" problem solver, able to consider and weight variables such as speed, cost, and personal preferences, and allow personalization based on repeated journeys, as well as link to calendar and scheduling data, and interactive prompts.

Types of AI: S — SA — DL — GAN

SOCIAL MEDIA
SPEECH OR CONTENT MODERATION

Systems can facilitate human teams in identifying, flagging, and deleting posts with defined, prohibited terms (such as "hate speech" or profanity). Categorizing and selectively reacting based on platform policies, usually embedded in human/computer systems for review and decision.

Value of AI: More efficient at scale than human-alone reviews. Additionally, well-designed systems can potentially adapt to variations in context, intent, cultural norms, and user expectations more consistently across platforms.

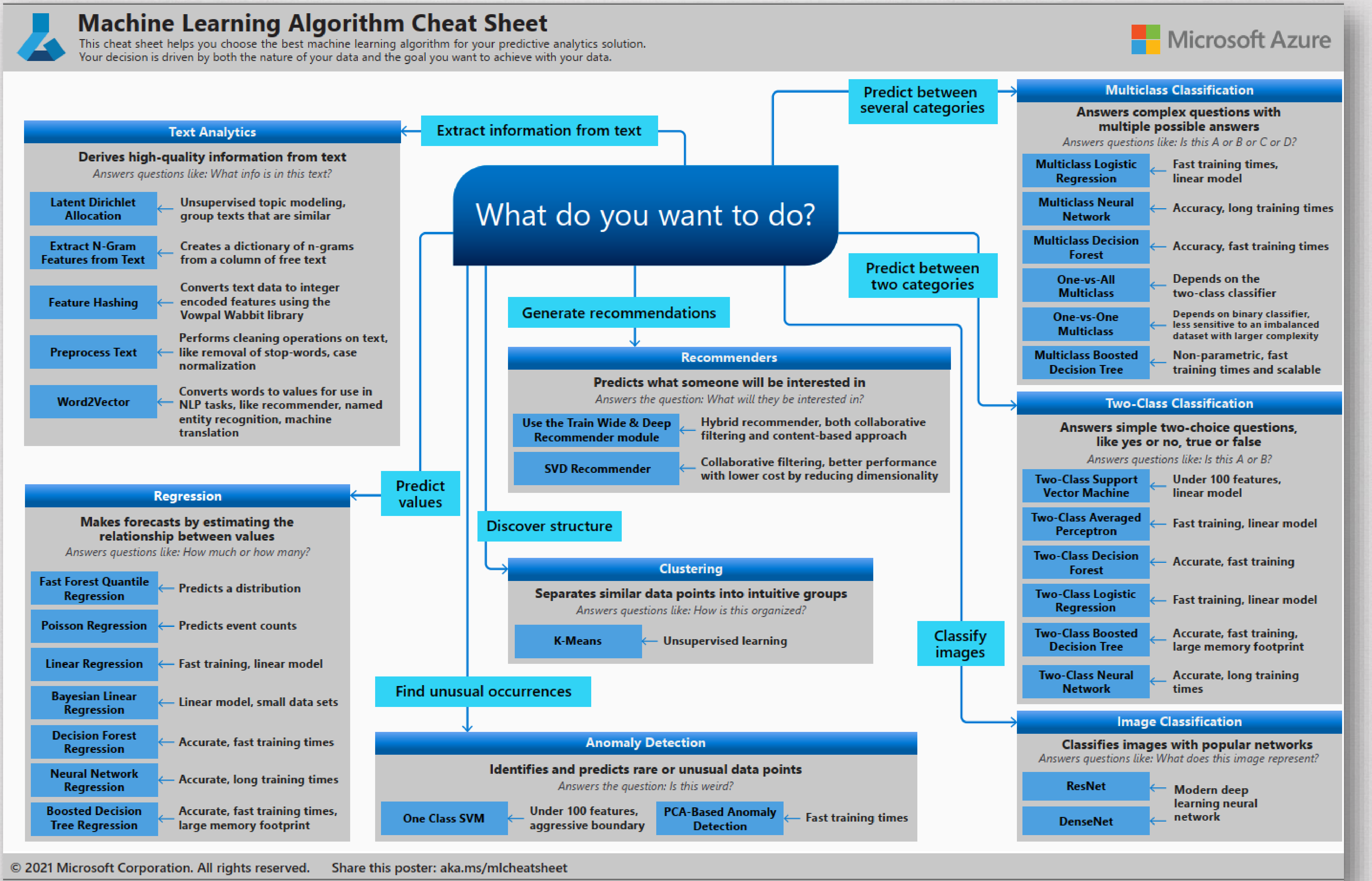
Types of AI: KE — NLP — RL

FORECASTING
SUPPLY CHAIN MANAGEMENT

Systems to improve traditional inventory and forecasting beyond historical/internal trend data, to weight and include external factors such as weather, consumer sentiment, demographic trends, analysis of portal traffic, stock fluctuations, and service levels

Value of AI: Systems can increase accuracy and efficiency, as well as provide improved transparency and reliable, predictive analytics; enable aggregate forecasting from individual impact up through regional levels.

Types of AI: P&S — ES — KE — ML



Machine Learning Algorithm Cheat Sheet

This cheat sheet helps you choose the best machine learning algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the goal you want to achieve with your data.

Microsoft Azure

What do you want to do?

Extract information from text → Text Analytics

Generate recommendations → Recommenders

Predict values → Regression

Discover structure → Clustering

Find unusual occurrences → Anomaly Detection

Predict between several categories → Multiclass Classification

Predict between two categories → Two-Class Classification

Classify images → Image Classification

Text Analytics

Derives high-quality information from text

Answers questions like: What info is in this text?

Latent Dirichlet Allocation	Unsupervised topic modeling, group texts that are similar
Extract N-Gram Features from Text	Creates a dictionary of n-grams from a column of free text
Feature Hashing	Converts text data to integer encoded features using the Vowpal Wabbit library
Preprocess Text	Performs cleaning operations on text, like removal of stop-words, case normalization
Word2Vector	Converts words to values for use in NLP tasks, like recommender, named entity recognition, machine translation

Regression

Makes forecasts by estimating the relationship between values

Answers questions like: How much or how many?

Fast Forest Quantile Regression	Predicts a distribution
Poisson Regression	Predicts event counts
Linear Regression	Fast training, linear model
Bayesian Linear Regression	Linear model, small data sets
Decision Forest Regression	Accurate, fast training times
Neural Network Regression	Accurate, long training times
Boosted Decision Tree Regression	Accurate, fast training times, large memory footprint

Recommenders

Predicts what someone will be interested in

Answers the question: What will they be interested in?

Use the Train Wide & Deep Recommender module	Hybrid recommender, both collaborative filtering and content-based approach
SVD Recommender	Collaborative filtering, better performance with lower cost by reducing dimensionality

Clustering

Separates similar data points into intuitive groups

Answers questions like: How is this organized?

K-Means	Unsupervised learning

Anomaly Detection

Identifies and predicts rare or unusual data points

Answers the question: Is this weird?

One Class SVM	Under 100 features, aggressive boundary
PCA-Based Anomaly Detection	Fast training times

Multiclass Classification

Answers complex questions with multiple possible answers

Answers questions like: Is this A or B or C or D?

Multiclass Logistic Regression	Fast training times, linear model
Multiclass Neural Network	Accuracy, long training times
Multiclass Decision Forest	Accuracy, fast training times
One-vs-All Multiclass	Depends on the two-class classifier
One-vs-One Multiclass	Depends on binary classifier, less sensitive to an imbalanced dataset with larger complexity
Multiclass Boosted Decision Tree	Non-parametric, fast training times and scalable

Two-Class Classification

Answers simple two-choice questions, like yes or no, true or false

Answers questions like: Is this A or B?

Two-Class Support Vector Machine	Under 100 features, linear model
Two-Class Averaged Perceptron	Fast training, linear model
Two-Class Decision Forest	Accurate, fast training
Two-Class Logistic Regression	Fast training, linear model
Two-Class Boosted Decision Tree	Accurate, fast training, large memory footprint
Two-Class Neural Network	Accurate, long training times

Image Classification

Classifies images with popular networks

Answers questions like: What does this image represent?

ResNet	Modern deep learning neural network
DenseNet	Modern deep learning neural network

Share this poster: aka.ms/mlcheatsheet

RESOURCES

- NIST AI Risk Management Framework
  - NIST AI Risk Management Framework
- IAPP Glossary of Privacy Terms
  - Key Privacy Terms
- IAPP AI Governance Center
  - AI Governance Center (iapp.org)
  - Key AI Terms
- OECD.AI Policy Observatory
  - AI-Principles Overview - OECD.AI
  - OECD's live repository of AI strategies & policies - OECD.AI
- Partnership on AI
  - Home - Partnership on AI
- AI Incident Database
  - Welcome to the Artificial Intelligence Incident Database
- Future of Privacy Forum
  - AI & Machine Learning – Future of Privacy Forum (fpf.org)
  - Generative AI Checklist (Print Version) (fpf.org)
- Privacy Library of Threats 4 Artificial Intelligence
  - PLOT 4 AI
- LinkedIn Learning Courses
  - Introduction to Artificial Intelligence (1 hr 34 min)
  - Artificial Intelligence Foundations: Thinking Machines (1 hr 27 min)
  - Artificial Intelligence Foundations: Machine Learning (1 hr 50 min)
  - What is Generative AI (42 min)
  - AI Trends (1 hr 15 min)
  - Over 4000 Courses covering engineering aspects, Chat GPT focused, etc.
- Microsoft AI
  - Machine Learning Algorithm Cheat Sheet - designer - Azure Machine Learning | Microsoft Learn
  - Microsoft Responsible AI
- PWC Responsible AI Toolkit
  - PWC Responsible AI Toolkit
  - PWC AI Hub
- Salesforce AI Ethics & Maturity Model
  - Ethical AI Practices
  - Ethical AI Maturity Model

ADDITIONAL QUESTIONS OR THOUGHTS?

Contact:

Gabrielle Harris
CES Privacy Officer
Gabrielle_harris@byu.edu