Capturing and Preserving the AI process as paradata for accountability and audit-trail purposes

Dr. Patricia C. Franks
Professor Emerita
San Jose State University
patricia.franks@sjsu.edu

Paradata …

is an approach for documenting the AI process, which draws on multiple fields including empirical social sciences, XAI, and archival studies.

~Franks, Hamidzadeh, Cameron, ItrustAI Literature Review, "Positioning Paradata: documenting AI processes in recordkeeping and archives"

What do we mean by AI?

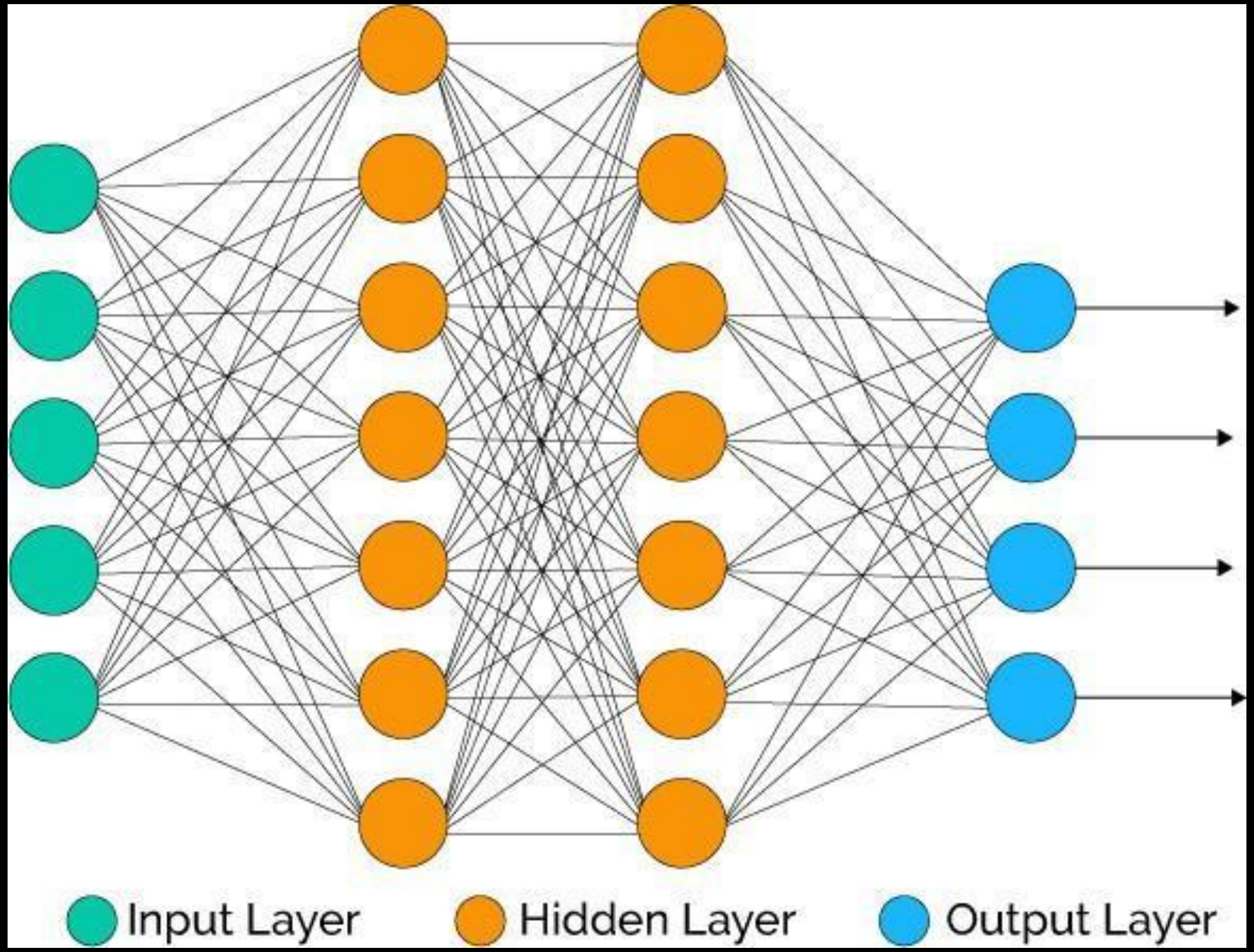
AI – Use Cases – Risks

How would you define AI?

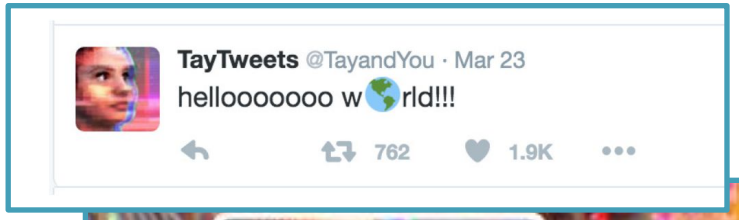
"Artificial Intelligence is software that can anticipate how a human would act, and then perform that action. It can learn to be more precise in its decision-making the more data it has, and through the algorithms it deploys."

~Interview with Elizabeth Perkes, Utah Department of Government Operations, Division of Archives and Records Service, Electronic Records Archivist 8/23/2022.


AI



What happens if the Output is not what was desired or expected?



It took less than 16 hours for Tay to begin spouting misogynistic and racist remarks

Microsoft is deleting its AI chatbot's incredibly racist tweets https://t.co/DUbL6M7WYg #TayTweets pic.twitter.com/TisQW4BqO7

‘ ?a???a? wac???e???er (@mattiaswac) March 24, 2016

AI applications can trigger serious social harms

The New York Times

Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men

Facebook called it "an unacceptable error." The co
struggled with other issues related to race.

Many Facial-Recognition Systems Are Biased, Says U.S. Study

Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

The New York Times

Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

So, where is the harm?

What are the consequences?

Who can/should be held accountable?

Microsoft's Tay: AI Tool: Natural Language Processing

- Harm: Tool manipulated to adopt offensive language
- Consequences: Offended Twitter users; embarrassed developer (Microsoft)
- Remediation: Pulled defective AI; Developed more "politically correct" replacement

Police Investigation: AI Tool: Facial Recognition

- Harm: Facial recognition output was considered infallible.
- Consequences: Arrest of Innocent Party; Distress to individual and family; loss of trust in police
- Remediation: Lawsuit against city of Detroit, police chief, and police detective; at a minimum re-evaluation of facial recognition software used.

ALEX DAVIES TRANSPORTATION FEB 29, 2016 2:04 PM

Google's Self-Driving Car Caused Its First Crash

Google's self-driving car appears to have caused its first crash on February 14, when it changed lanes and put itself in the path of an oncoming bus.

First-ever self-driving vehicle crash report released. Nearly all the WA wrecks involved Teslas

June 15, 2022 at 4:13 pm | *Updated June 15, 2022 at 5:51 pm*

My opinion is it's a bridge too far to go to fully autonomous cars."
~Elon Musk, Businessman, 2013

Who should be held accountable? The manufacturer? The driver? Both? Neither?

Elon Musk says Tesla will have self-driving cars without the need for human drivers this time next year

Fred Lambert - May. 22nd 2022 10:52 am PT @FredericLambert

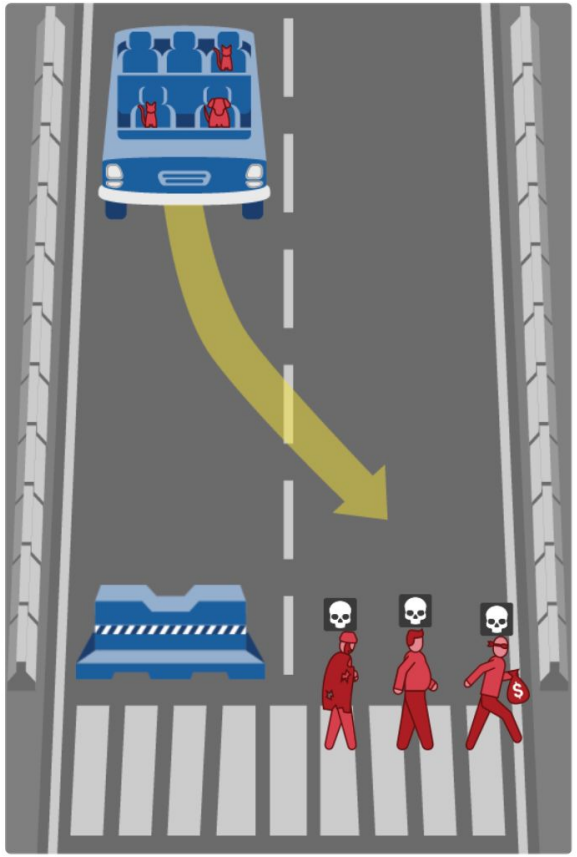
MORAL MACHINE

What should the self-driving car do?

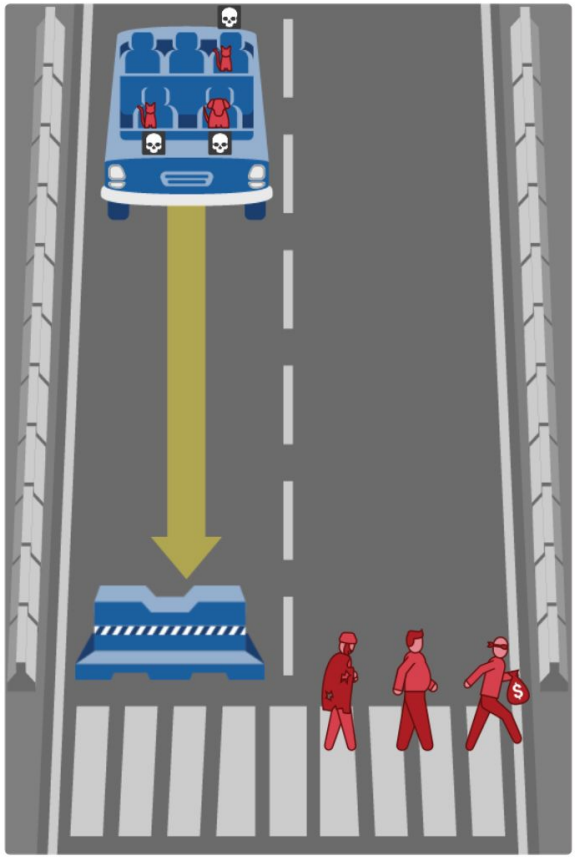
In this case, the self-driving car with sudden brake failure will swerve and drive through a pedestrian crossing in the other lane. This will result in ...

Dead:

- 1 homeless person
- 1 large man
- 1 criminal



Hide Description



Hide Description

2 / 13

In this case, the self-driving car with sudden brake failure will continue ahead and crash into a concrete barrier. This will result in ...

Dead:

- 1 dog
- 2 cats

https://www.moralmachine.net/

Toward Accountability & Auditable AI

Risks – Ethics – Guidance - Standards – Best Practices

Clearview AI's Facial Recognition Platform Achieves Superior Accuracy & Reliability Across All Demographics in NIST Testing

Clearview's algorithm ranks No. 1 in the U.S. in all categories as verified by National Institute of Standards & Technology (NIST) Facial Recognition Vendor Test (FRVT)

It ranked No. 1 in the U.S. for its performance in matching

- VISA Photos (99.81 percent)
- MUGSHOT Photos (99.76 percent)
- VISABORDER photos (99.7 percent) and
- BORDER Photos (99.42 percent)

It also ranked in the top five worldwide in all of these categories out of 650 algorithms.

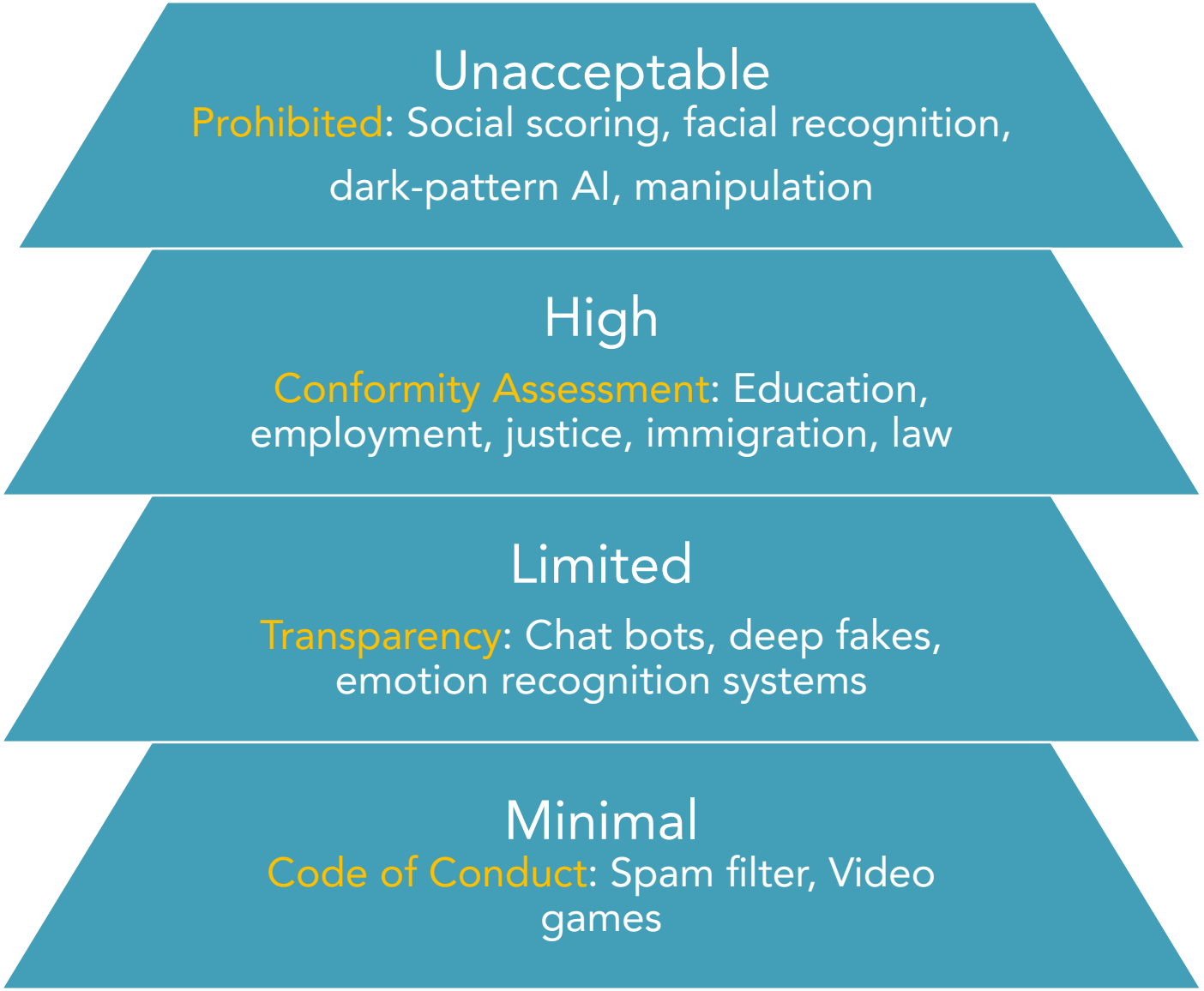
CLEARVIEW AI 2.0

Accelerate Cases with Publicly Available Facial Images

Online OSINT Images You Won't Find Any Other Way

20+ Billion Facial Images & Customizable Galleries

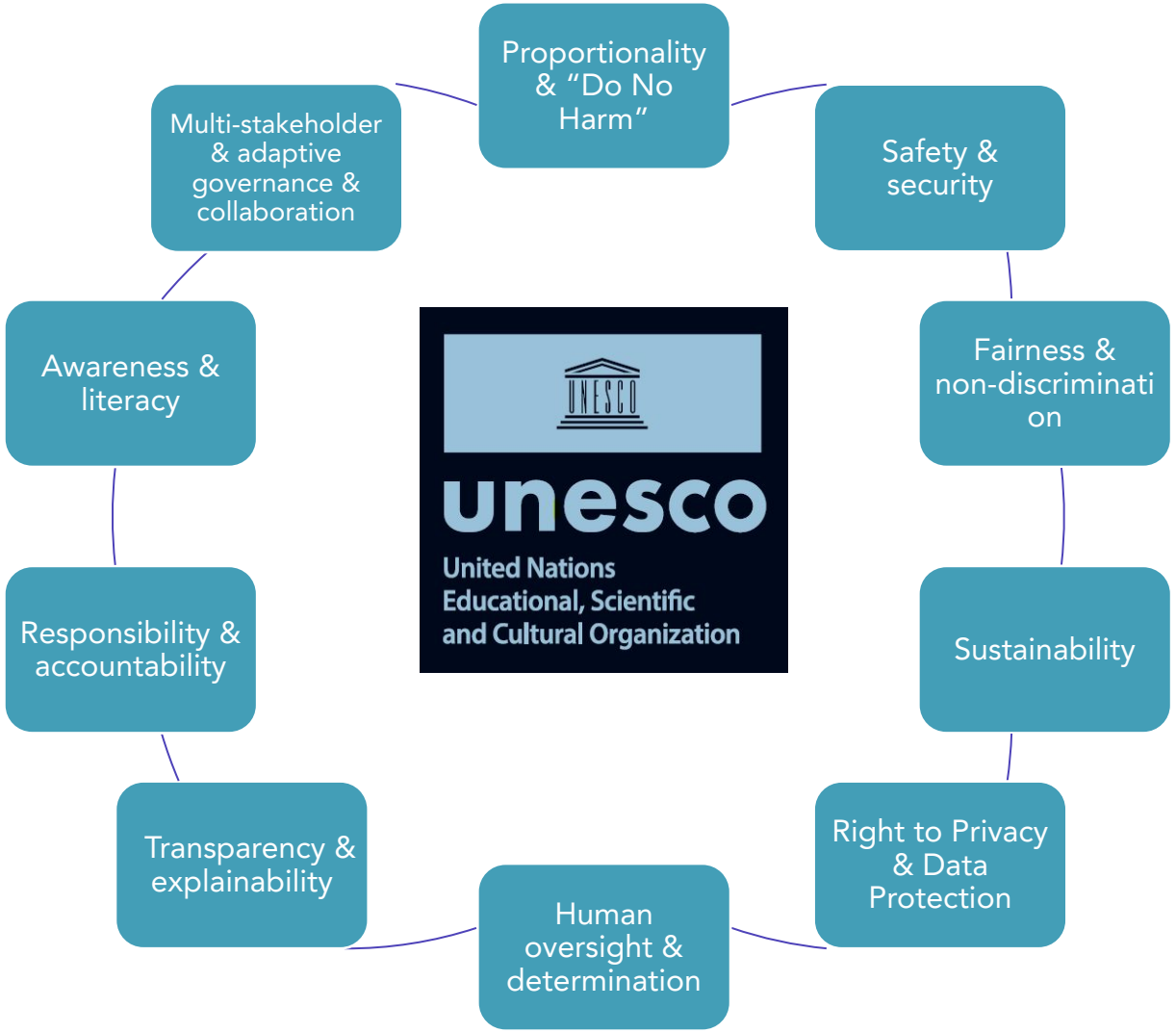
A Layered Risk-based Approach to AI Implementation



Based on the EU proposed Regulation on Artificial Intelligence (the EU AI Act) likely to be passed into law the first half of 2023.

https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

Human-centered AI – for the greater interest of the people--and not the other way around.



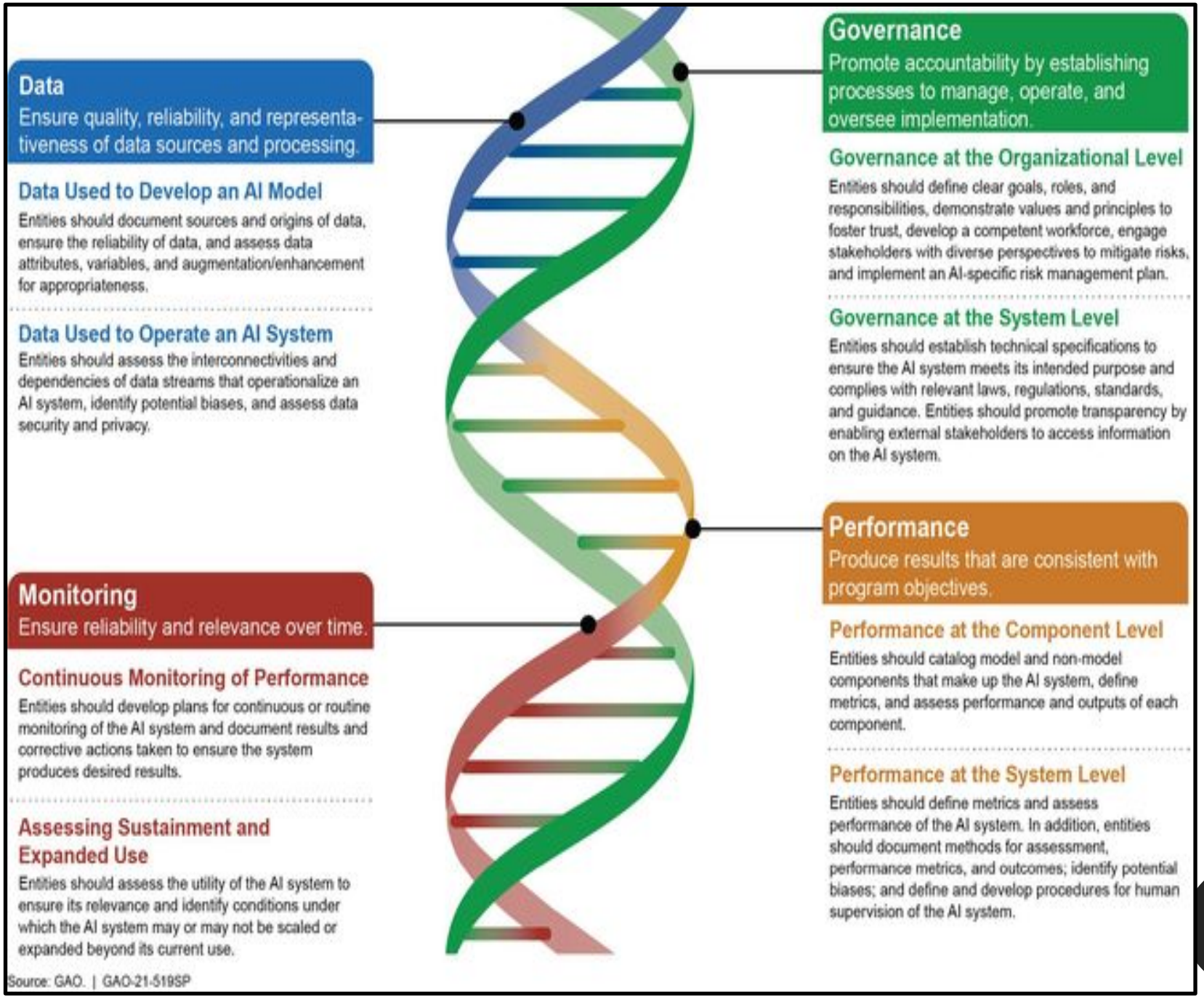
UNESCO plan for "ethical AI" was adopted in 2021.

https://unesdoc.unesco.org/ark:/48223/pf0000377897

AI Accountability Framework

U.S. Government Accountability Office (2021)

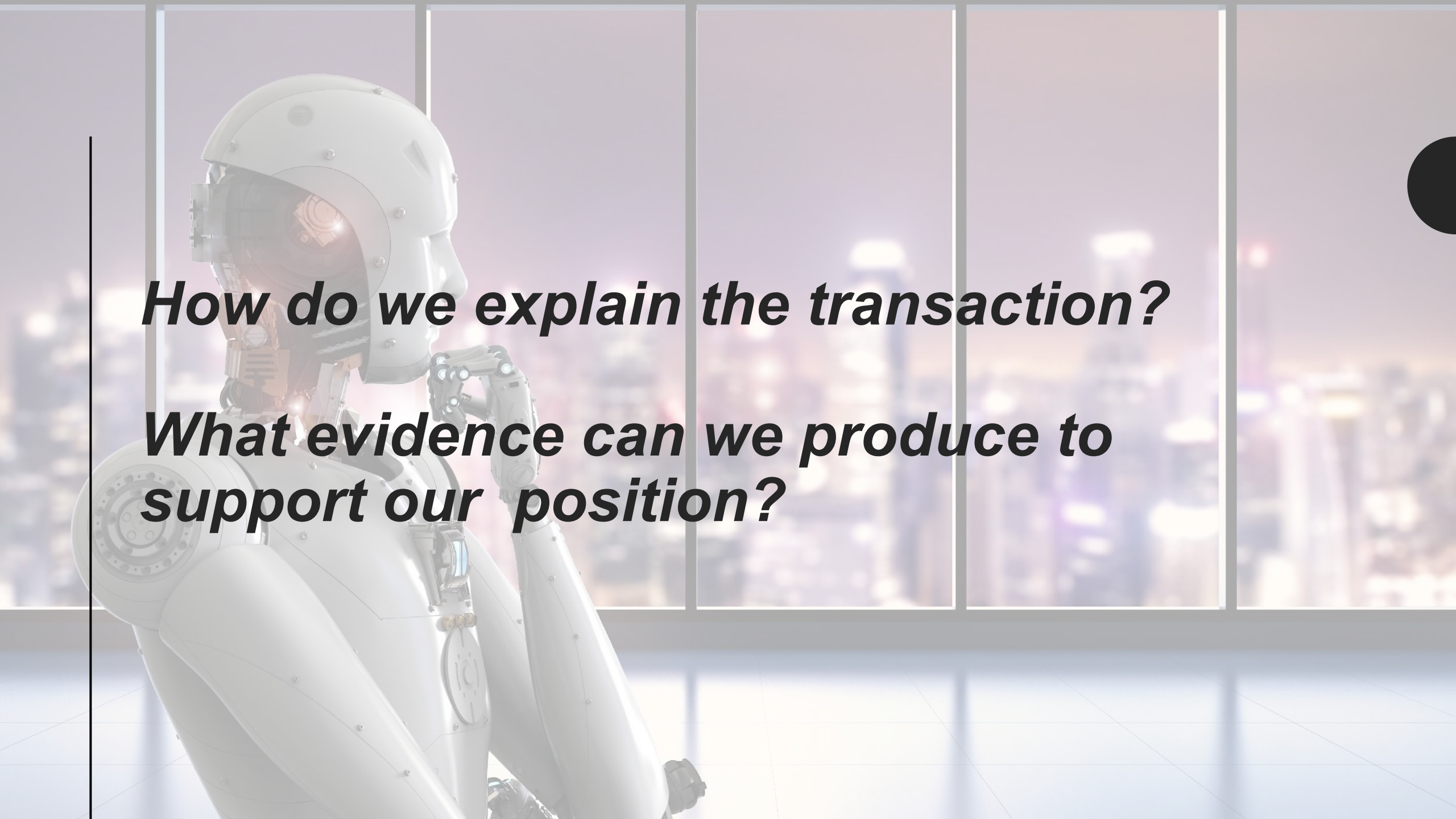
https://www.gao.gov/products/gao-21-519sp



What do we mean by Audit Trail & Auditable AI?

- An audit trail is the documented flow of a transaction. It is a detailed, chronological record whereby project details are tracked and traced.
- It should include information to establish what events occurred and who (or what) caused them).
- Beyond responsible/accountable AI, is Auditable AI, an audit trail of a company's documented development governance standard during the production of an AI model.

https://www.fico.com/blogs/beyond-responsible-ai-8-steps-auditable-artificial-intelligence



How do we explain the transaction?

What evidence can we produce to support our position?

Documenting the AI Process

Documentation – Records - Evidence

"Defining an AI record and developing methods for capturing AI records is a project the profession should take on."

~Norman Mooradian, Ph.D.

“If business is no longer to be transacted only by human beings, but also by AI agents, or some combination of the two, what will evidence of those transactions look like, what will the record be?”

~Jenny Bunn

Recordkeepers may ask:

What records are created within AI research teams to document their process?

What records are created of the decisions to procure or deploy systems utilising AI?

What records are created of the decisions and impact of such systems?

Are the created records sufficient to meet existing legal provisions?

Do the created records meet the required standards of quality?

~Jenny Bunn, 2020, p. 14

Paradata is a source of information in the form of auxiliary data describing the process [of the use of computer-assisted survey instruments.] *~Mick P. Couper, 2010*

This brings us back to

Paradata

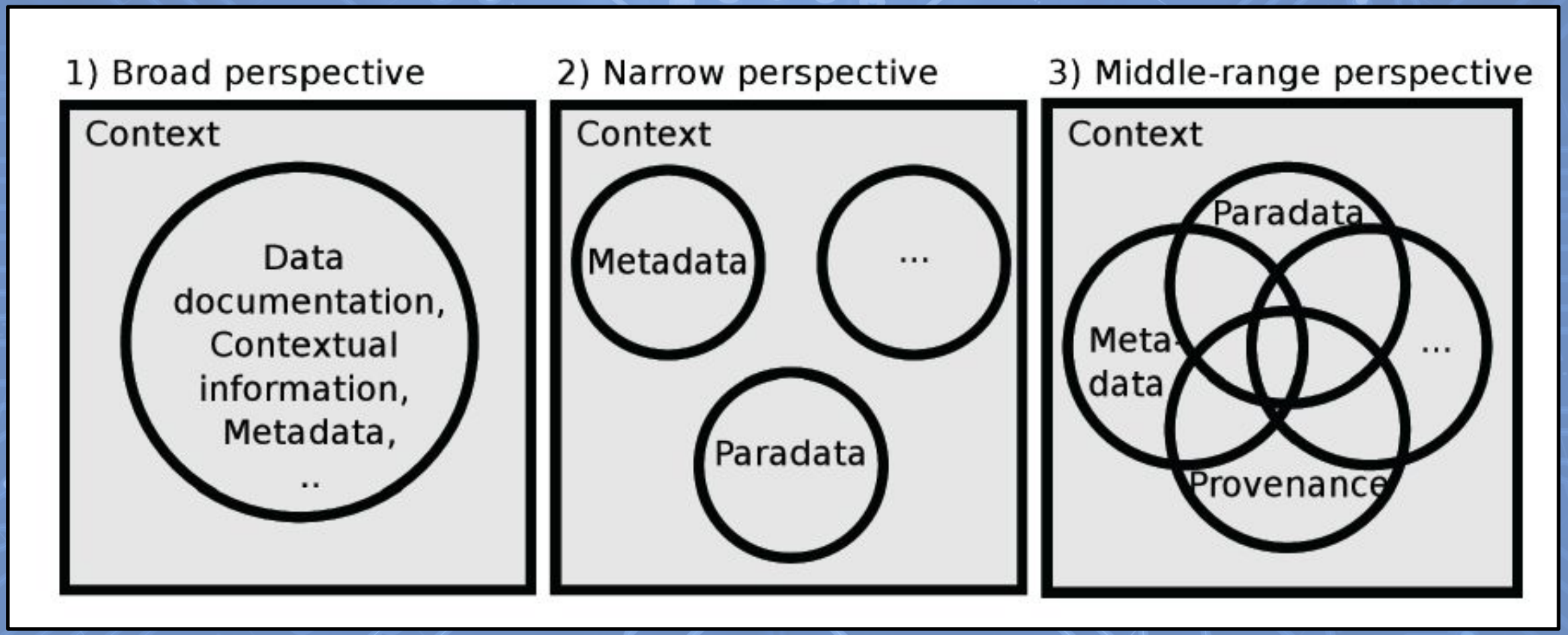
Paradata is a term used to describe data generated as a by-product of the data collection process. *~U.S. Census Bureau, 2022*

Metadata is formalized data about statistical data needed to search for, display, and analyze those data. ~*National Institute of Statistical Sciences, 2010*

Metadata or Paradata?

Paradata is formalized data on methodologies, processes and quality associated with the production and assembly of statistical data. ~*National Institute of Statistical Sciences, 2010*

Perspectives to Contextual Information

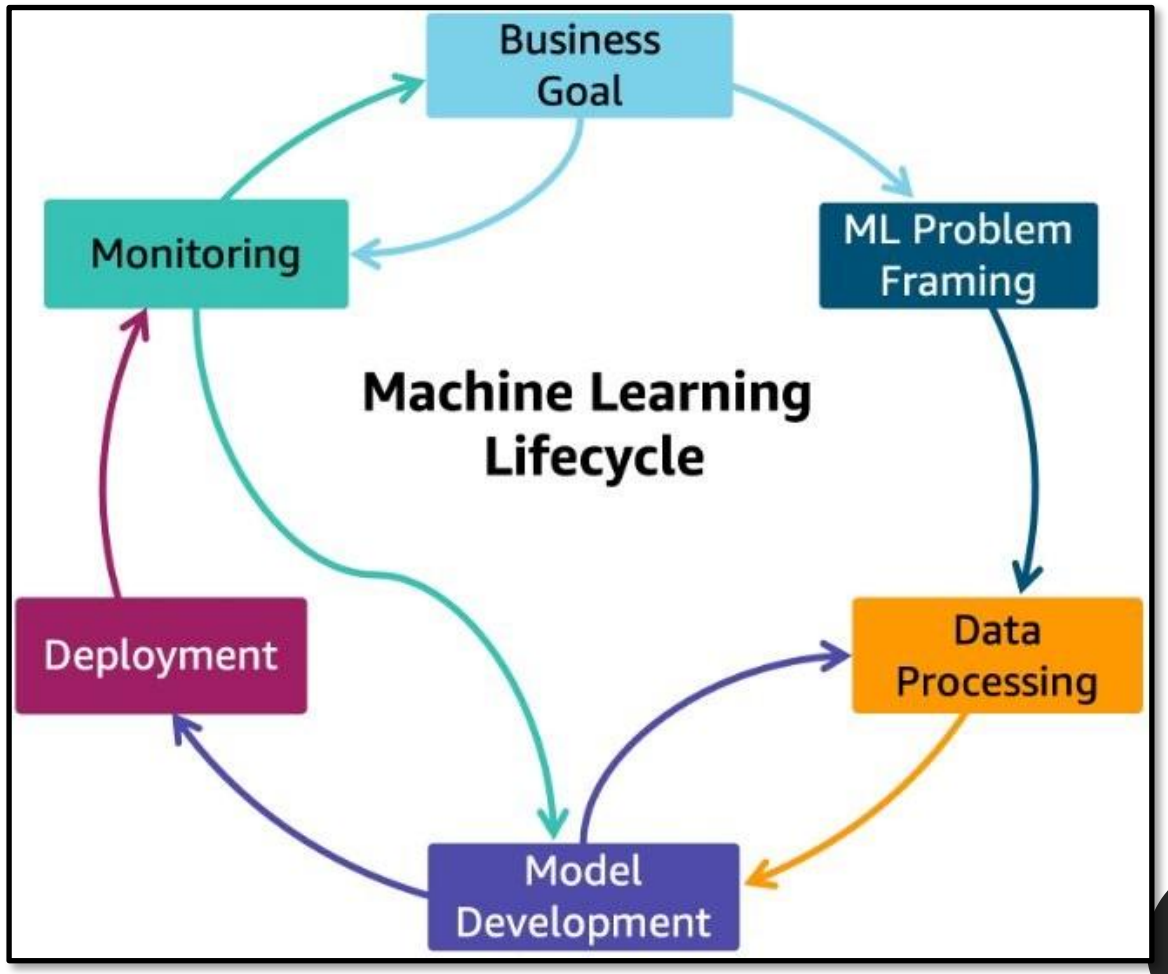


~Huvila, Isto, "Improving the usefulness of research data with better Paradata"

PARADATA & AI Process

Paradata is the formation about the procedure(s) and tools used to create and process information resources, along with information about the persons carrying out those procedures.

~ITrustAI working definition



Data Science & AI Lifecycle - a General View

Scope

Requirements

1. **Explore** multiple business ideas
2. **Assess Feasibility**
3. **Prioritize and Select** Use Cases...
4. **Establish Business KPIs** for use cases
5. **Establish swim lanes** for each use case

Business User

Data Scientist

Understand

Data Assets

1. **All: Use** Heterogeneous Sources, Projects, General Policies, etc.
2. *Provider:* **Discover**, Track Lineage, Catalog, Curate, Classify, Grant Access
3. *Steward:* Create **Policies** for Data Assets, Grant Access
4. *Consumer*: Find, Understand, Add, Explore, Review
5. Make Explored Data Available for Collaboration

Data Steward
Data Provider
Data Consumer

Build (Dev)

ML/DL assets

Connect to Needed Data

1. **Analyze Data**
2. **Prepare Data**
3. **Build & Train Model**
   - Code, Canvas
4. **Evaluate Model**
5. **Collaborate; Tag for Review**
6. **Publish to Catalog** (models, notebooks)

Data Engineer
Data Scientist

Deploy and Run (QA and Prod)

Business KPIs

1. **Review for Deploy** after code review, third party oversight, unit tests **Tag Deployable Version**
2. **Deploy Model** to ML Runtime Engine
3. **Monitor & Evaluate** model execution
4. **Manage**
   - View vs. thresholds
   - Model versioning

Software Engineer

Data Scientist

Business User

Monitor and Manage (QA and Prod)

Business KPIs

Configure Model *for Monitoring & Integrations (new and updates)*

Obtain Model Insights

1. **Quality**
2. **Perf. (throughput)**
3. **Custom Metrics**
4. **Fairness**

5. **Explain Transactions** (on-demand)

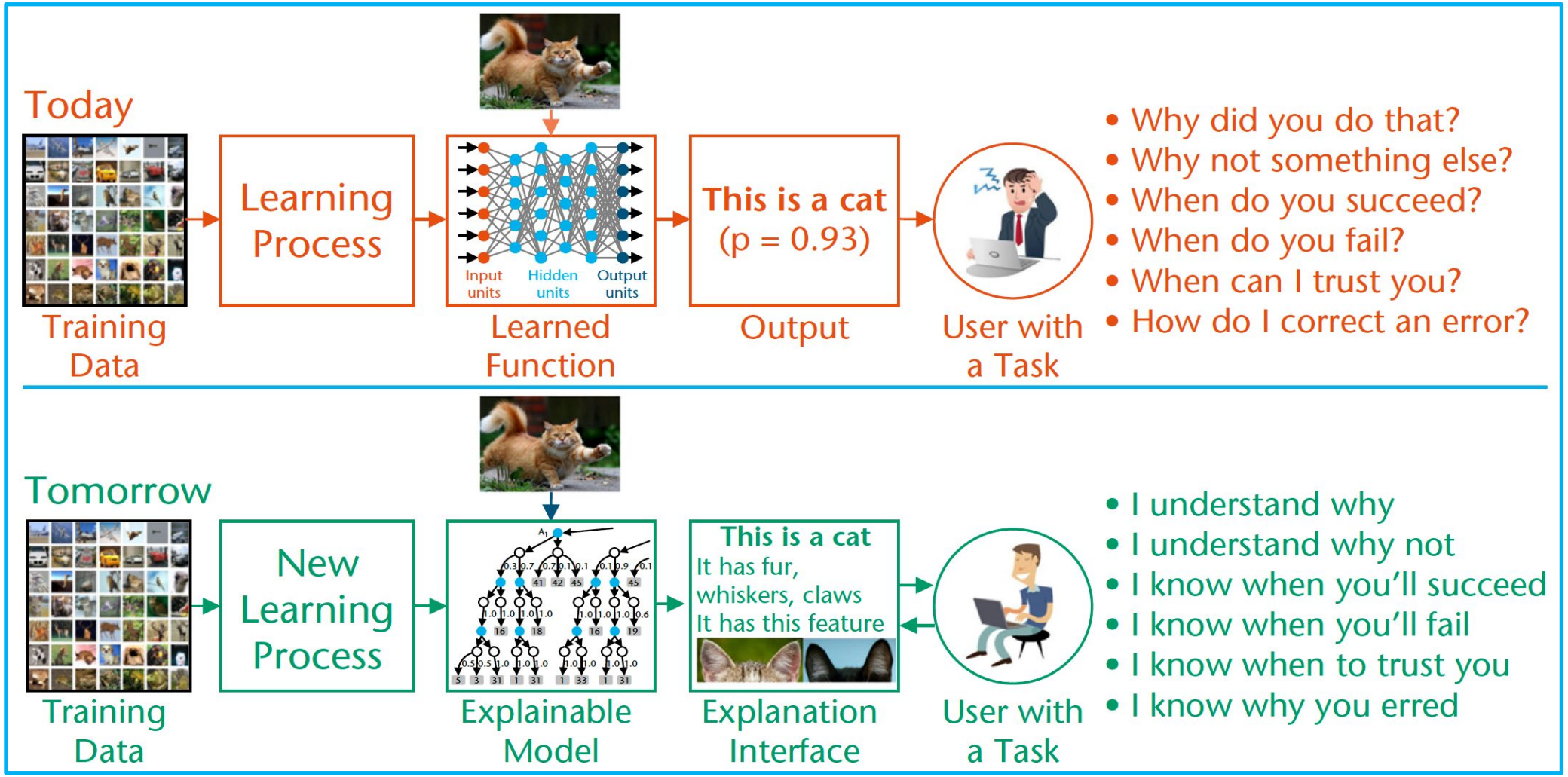
AI Operations

Business User

CI/CD Pipeline

IBM

XAI Concept



~Turek, DARPA

Modes of Explanation

- Causal – How it functions
- Epistemic – How we know it functions
- Justificatory – On what grounds it functions

Justificatory -- Can refer to AI system properties (e.g., datasets and algorithms).

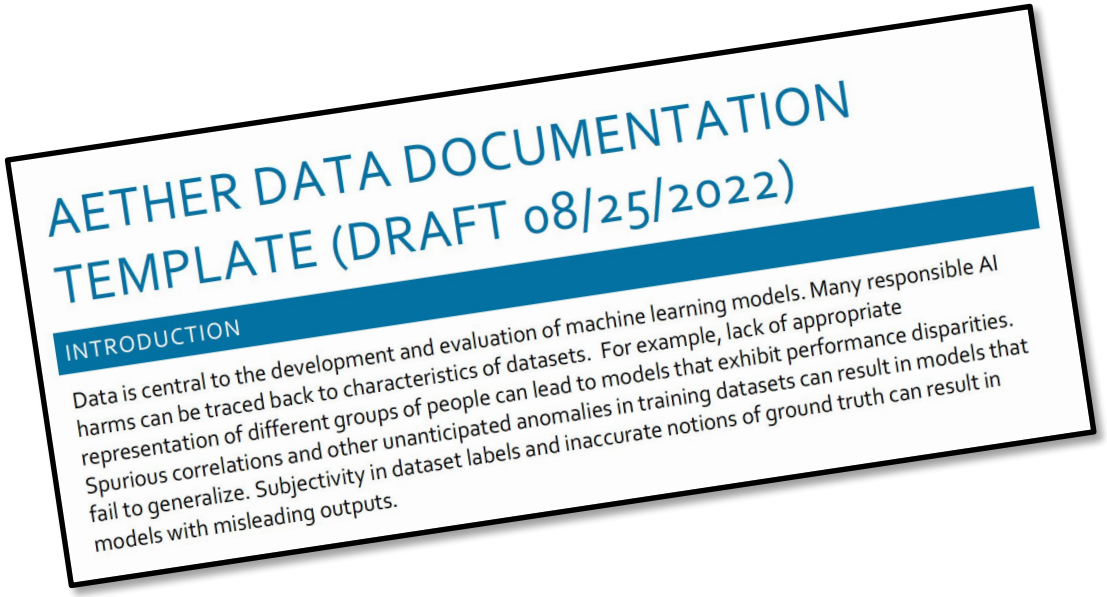
Must also reference institutional and social facts about the implementation of the system (e.g., regulations, standards, organizational processes pertinent to the use case).

ISO/IEC TR 24028 (2020-05) Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

Microsoft's Datasheets for Datasets

Potential Audience:

- People who are considering using this dataset to train or evaluate models
- People who are auditing a model or AI system



AETHER DATA DOCUMENTATION TEMPLATE (DRAFT 08/25/2022)

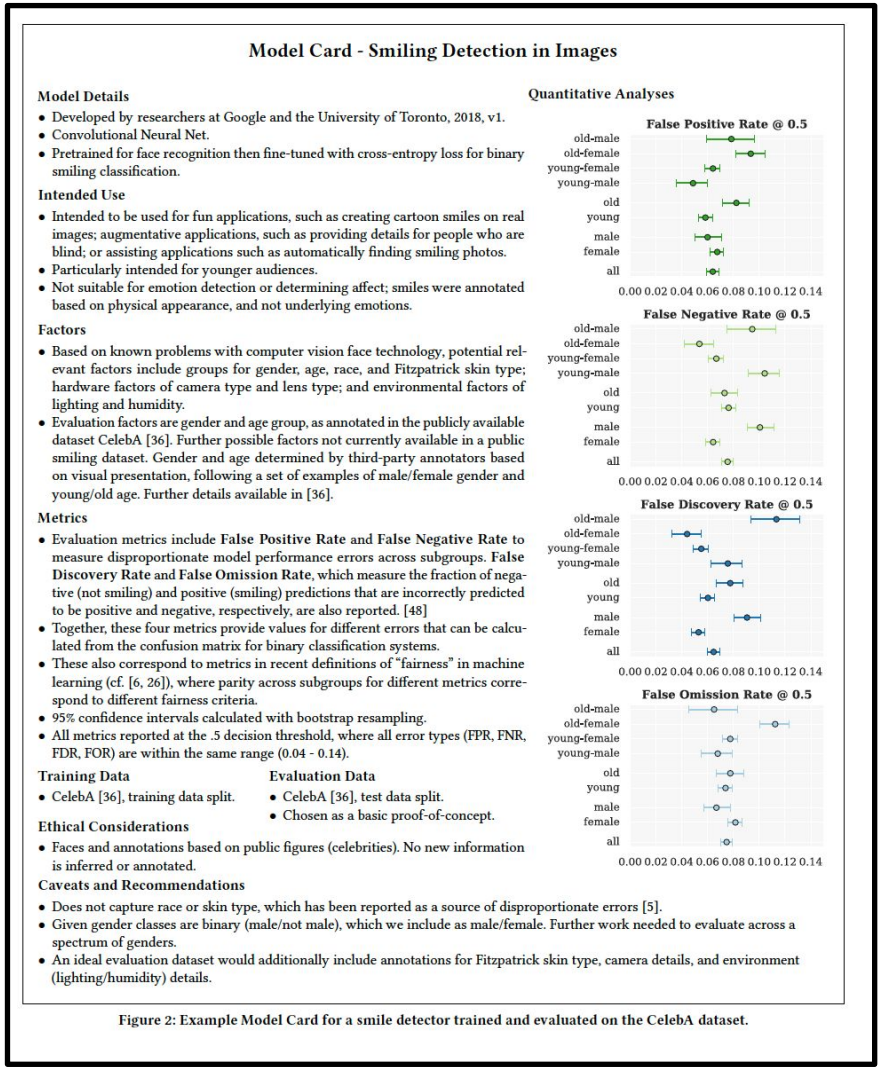
INTRODUCTION

Data is central to the development and evaluation of machine learning models. Many responsible AI harms can be traced back to characteristics of datasets. For example, lack of appropriate representation of different groups of people can lead to models that exhibit performance disparities. Spurious correlations and other unanticipated anomalies in training datasets can result in models that fail to generalize. Subjectivity in dataset labels and inaccurate notions of ground truth can result in models with misleading outputs.

Major sections of the template:

- Data Set Overview (ex. contact, distribution, access basics; data set contents; intended and inappropriate uses.)
- Details (data collection procedures/ representativeness; data quality; pre-processing cleaning, and labeling; privacy; additional details on distribution and access.)

Google Model Cards



Model Card - Smiling Detection in Images

Model Details
- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use
- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors
- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics
- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data
- CelebA [36], training data split.

Evaluation Data
- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations
- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations
- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses

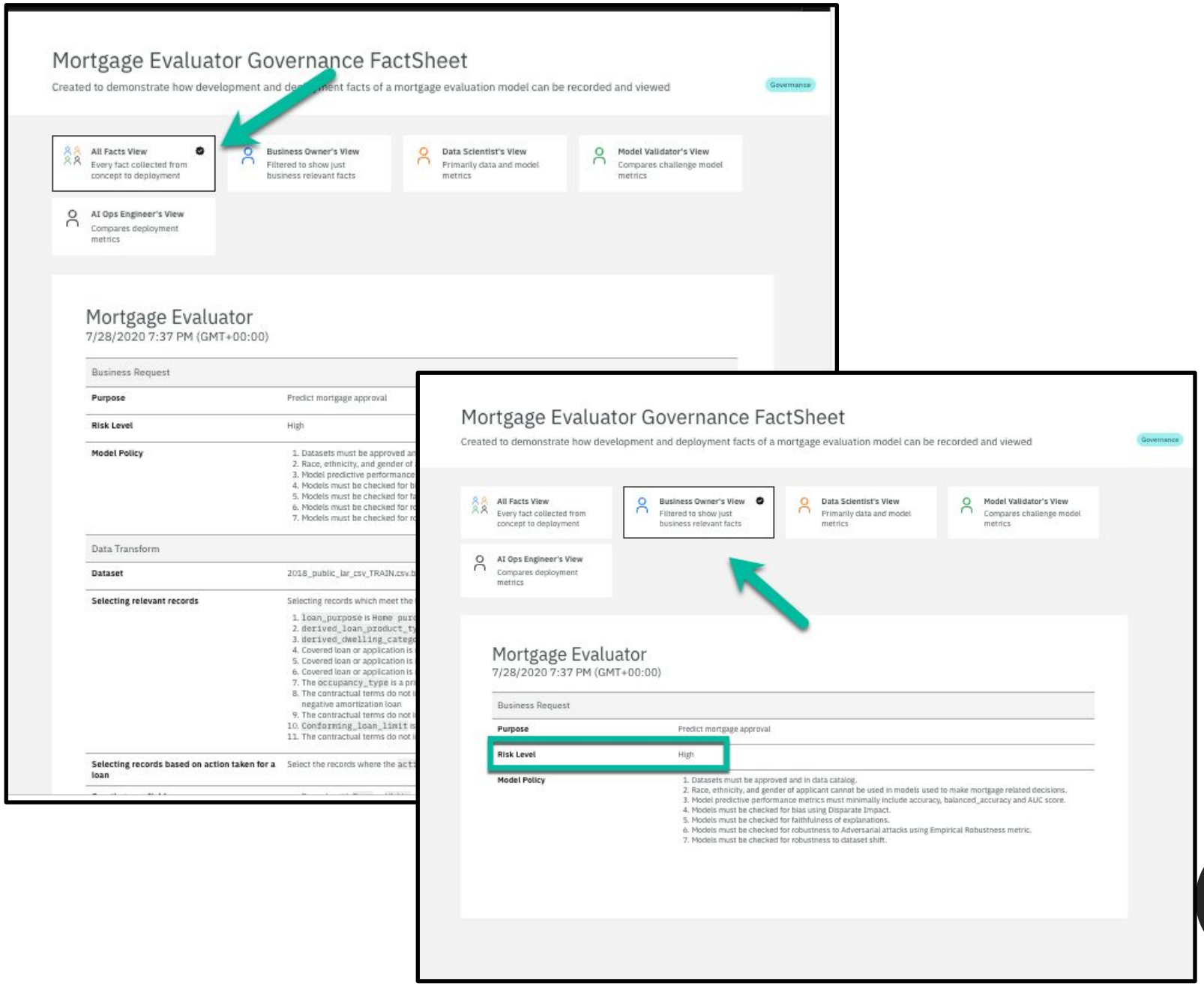
Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

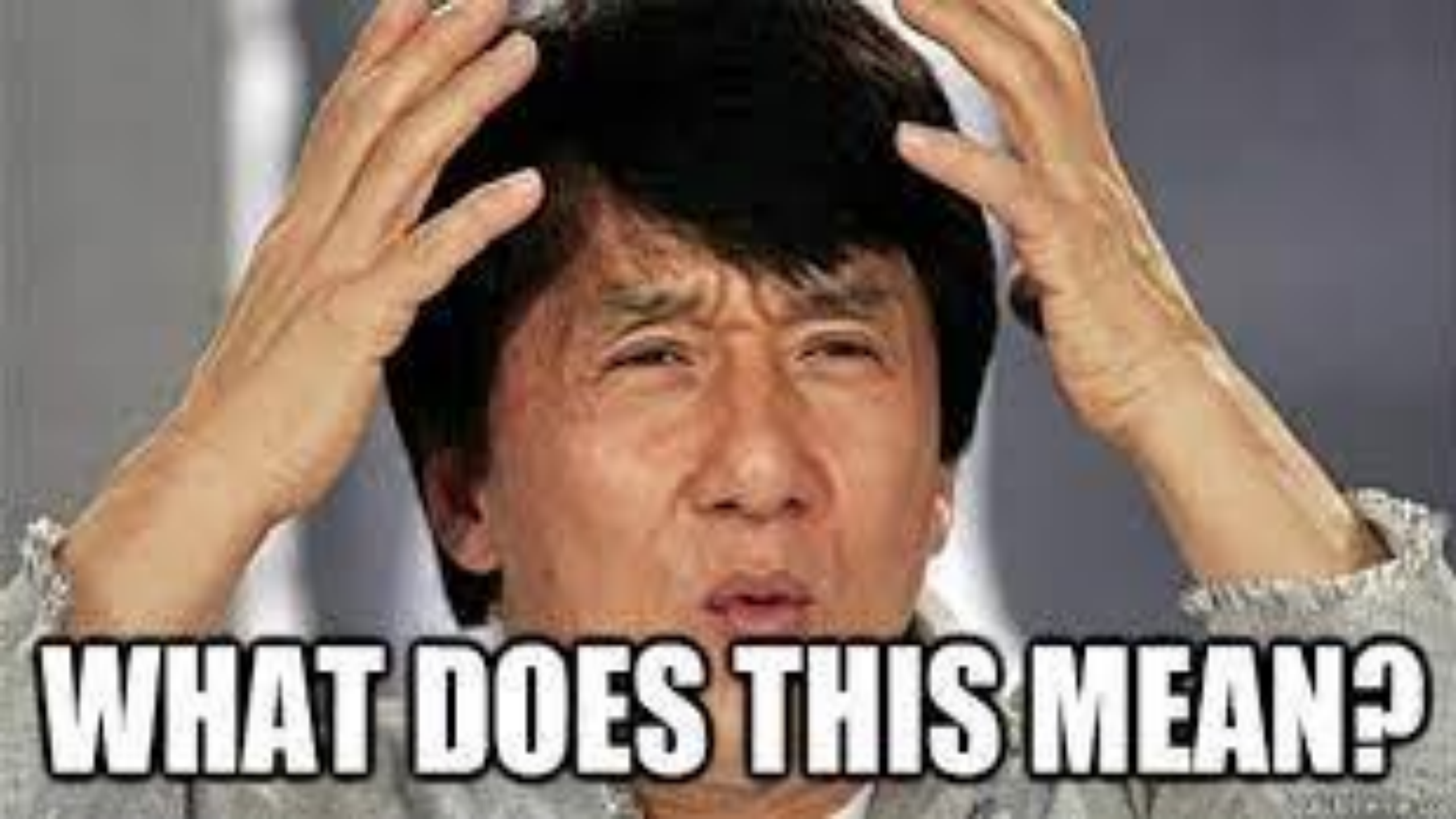
- ✔ Model Details
- ✔ Intended Use
- ✔ Factors
- ✔ Metrics
- ✔ Training Data
- ✔ Ethical Considerations
- ✔ Caveats & Recommendations

Updated model cards at https://modelcards.withgoogle.com/about

IBM's AI FactSheets 360

5 views – all facts and based on roles




WHAT DOES THIS MEAN?

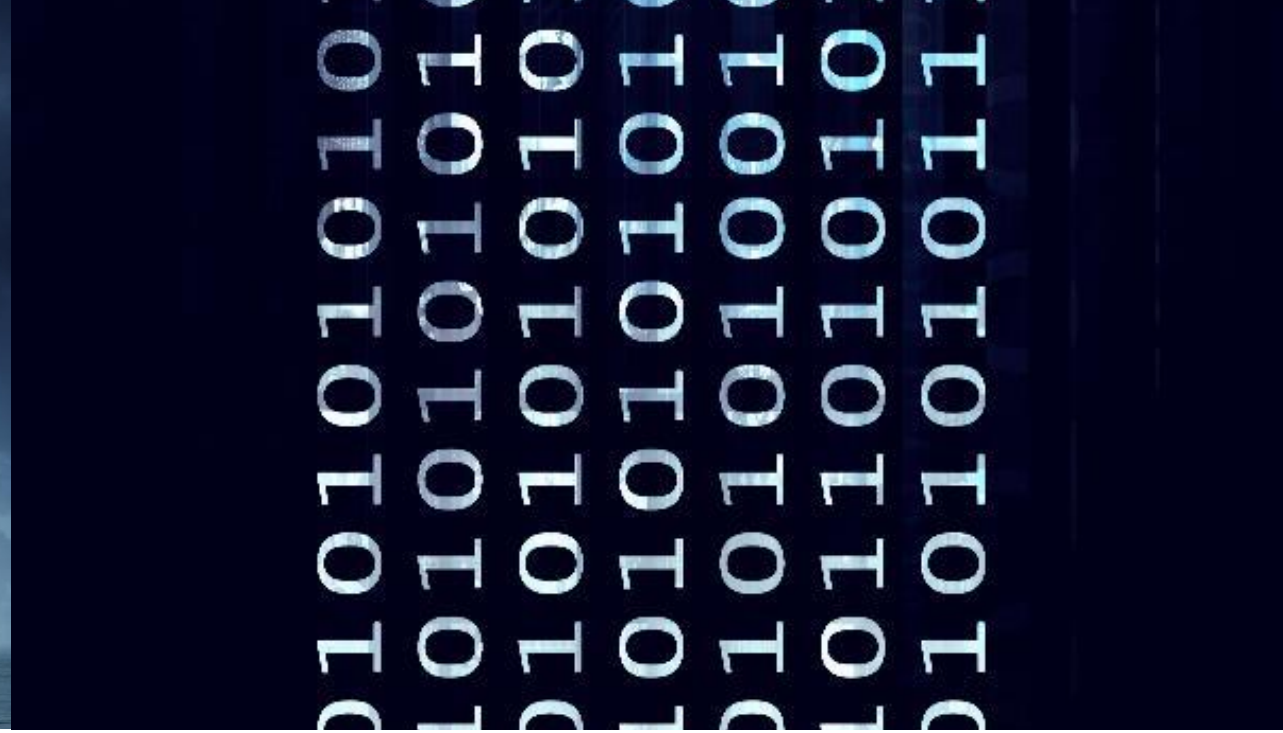
Interpretable implying some sense of understanding how the technology works;

Explainable, implying that a wider range of users can understand why or how a conclusion was reached;

Transparent implying some level of accessibility to the data or algorithm;

Justifiable implying there is an understanding of the case in support of a particular outcome

Contestable implying users have the information they need to argue against a decision or classification.



Paradata Can Help!

Continue research into…

- the nature of Paradata to document the AI process,
- the relationship between and potential overlap across Metadata and Paradata,
- the actions that take place along the AI lifecycle that require documentation,
- a recommended risk-management approach when determining the extent of documentation needed,
- the best form of representation, method of capture, and preservation,
- And finally, the identification and development of AI tools and techniques that can be employed to aid us in this task.

Thank you!

Patricia C. Franks, PhD
Professor Emerita
ItrustAI Researcher
San Jose State University
patricia.franks@sjsu.edu