Clear values, murky responsibilities:

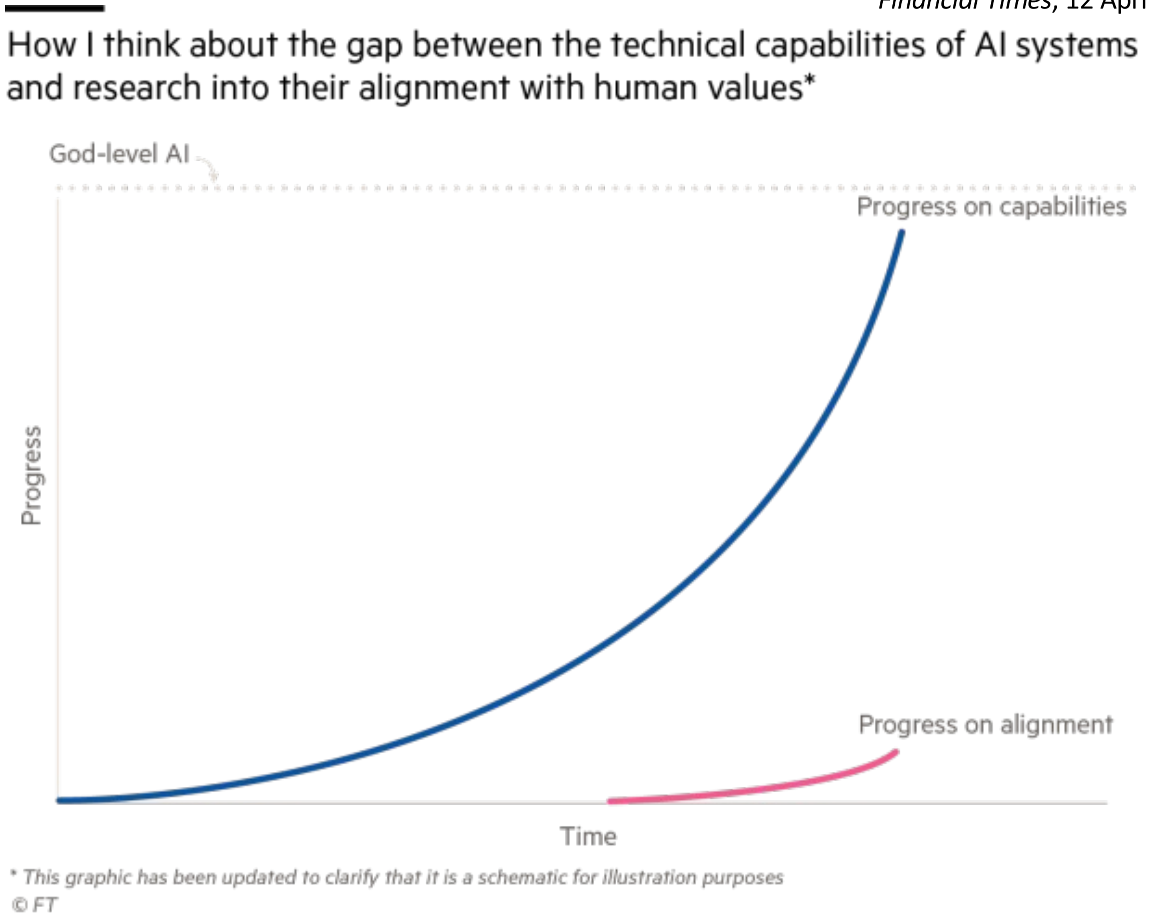
considering the ethical pipeline of archival information and AI implementation

Presented by
Jim Suderman
at the
InterPARES Summer School Symposium
San Benedetto, Italy,
7 July 2023

Aligning AI with Human Values

- Artificial General Intelligence (AGI): [definition]
  - a form of artificial intelligence that possesses the ability to understand, learn, and perform tasks across a wide range of domains at a level equal to or surpassing human capabilities
  - AGI, unlike narrow or specialized AI, can adapt, reason, and learn autonomously without being limited to a specific task or function.
    - Kareem Amer, "The Road to AGI: LLM and the Race Towards AGI" 2023

- Ian Hogarth, "We must slow down the race to God-like AI" *Financial Times*, 12 April 2023



How I think about the gap between the technical capabilities of AI systems and research into their alignment with human values*

* This graphic has been updated to clarify that it is a schematic for illustration purposes

© FT

Human Rights and Value to Society

“L'archivista si fa carico di difendere la professione per la sua utilità sociale, nel proteggere l'integrità e il valore probatorio degli archivi presta particolare attenzione a quelli che documentano diritti umani e libertà fondamentali...”

> [“The archivist takes it upon himself to defend the profession for its social utility, in protecting the integrity and probative value of archives pays particular attention to those that document human rights and fundamental freedoms…” translation by Google Translate.]

- Art. 3, CODICE DEONTOLOGICO, **Associazione Nazionale Archivistica Italiana** (2017)

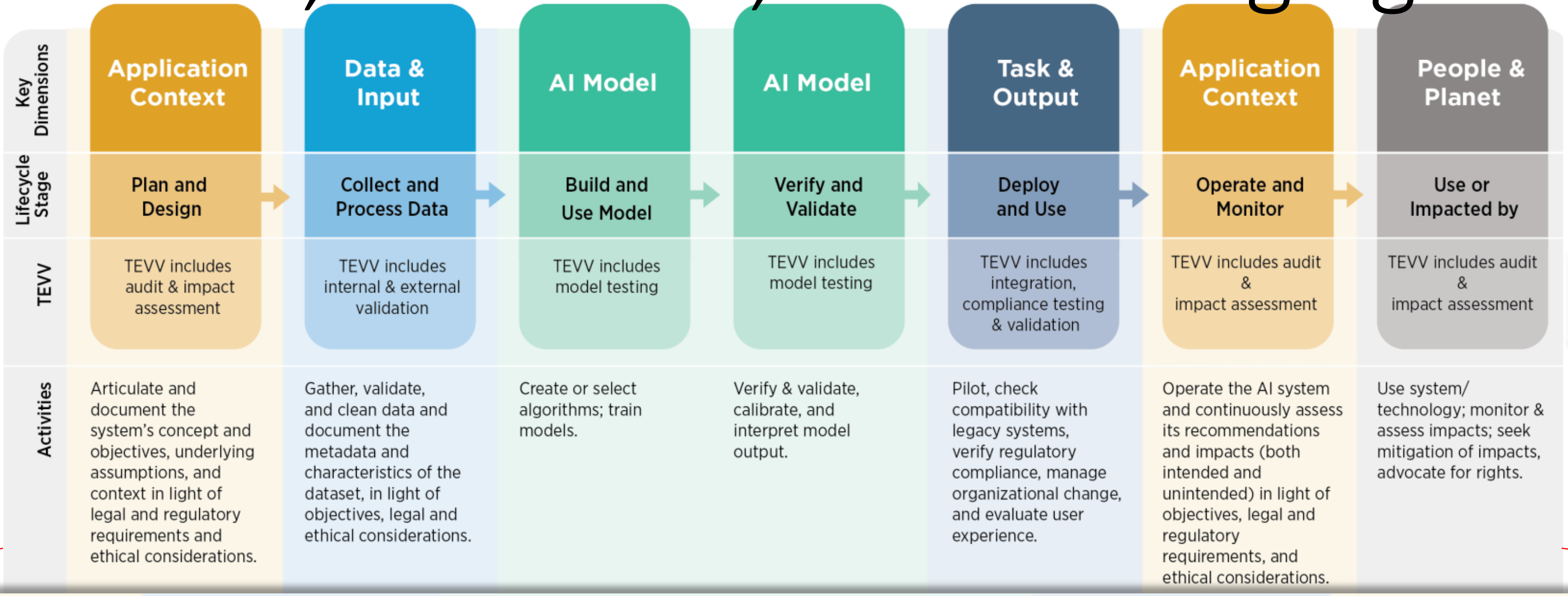
“Members should do everything within their power to avoid the destruction of documents that are of historical or public value, including value for documenting and/or redressing gross human rights violations and/or protecting human rights and fundamental freedoms for individuals and/or groups.”

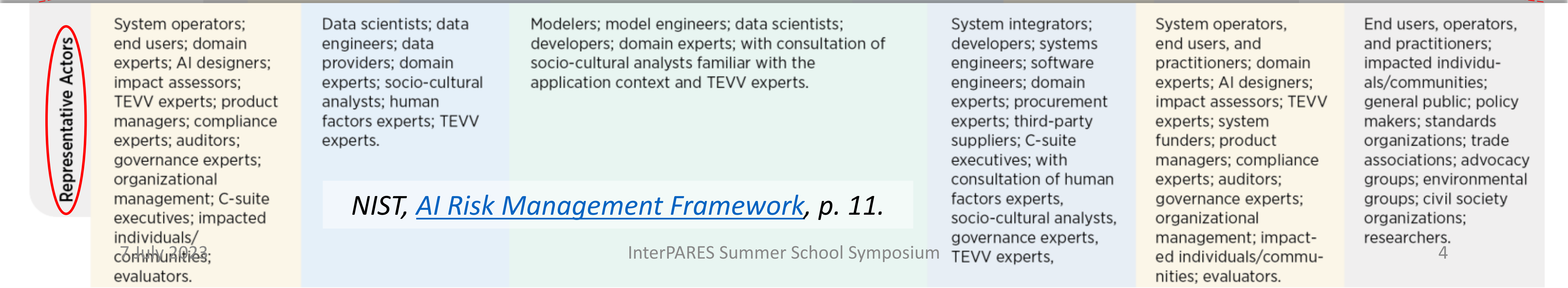
- Art. 11, Code of Ethics, **Archives and Records Association, UK and Ireland** (2020)

“These Guidelines articulate a framework for achieving Trustworthy AI based on fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (EU Charter), and in relevant international human rights law.”

- European Commission, **High-Level Expert Group on Artificial Intelligence**, “Ethics Guidelines for Trustworthy AI” (2019), p. 6.

Archivists, “AI Actors,” and Managing AI Risk





Key Dimensions	Application Context	Data & Input	AI Model	AI Model	Task & Output	Application Context	People & Planet
Lifecycle Stage	Plan and Design	Collect and Process Data	Build and Use Model	Verify and Validate	Deploy and Use	Operate and Monitor	Use or Impacted by
TEVV	TEVV includes audit & impact assessment	TEVV includes internal & external validation	TEVV includes model testing	TEVV includes model testing	TEVV includes integration, compliance testing & validation	TEVV includes audit & impact assessment	TEVV includes audit & impact assessment
Activities	Articulate and document the system's concept and objectives, underlying assumptions, and context in light of legal and regulatory requirements and ethical considerations.	Gather, validate, and clean data and document the metadata and characteristics of the dataset, in light of objectives, legal and ethical considerations.	Create or select algorithms; train models.	Verify & validate, calibrate, and interpret model output.	Pilot, check compatibility with legacy systems, verify regulatory compliance, manage organizational change, and evaluate user experience.	Operate the AI system and continuously assess its recommendations and impacts (both intended and unintended) in light of objectives, legal and regulatory requirements, and ethical considerations.	Use system/ technology; monitor & assess impacts; seek mitigation of impacts, advocate for rights.
Representative Actors	System operators; end users; domain experts; AI designers; impact assessors; TEVV experts; product managers; compliance experts; auditors; governance experts; organizational management; C-suite executives; impacted individuals/ communities; evaluators.	Data scientists; data engineers; data providers; domain experts; socio-cultural analysts; human factors experts; TEVV experts.	Modelers; model engineers; data scientists; developers; domain experts; with consultation of socio-cultural analysts familiar with the application context and TEVV experts.		System integrators; developers; systems engineers; software engineers; domain experts; procurement experts; third-party suppliers; C-suite executives; with consultation of human factors experts, socio-cultural analysts, governance experts, TEVV experts,	System operators, end users, and practitioners; domain experts; AI designers; impact assessors; TEVV experts; system funders; product managers; compliance experts; auditors; governance experts; organizational management; impacted individuals/communities; evaluators.	End users, operators, and practitioners; impacted individuals/communities; general public; policy makers; standards organizations; trade associations; advocacy groups; environmental groups; civil society organizations; researchers.

TEVV = test, evaluation, verification, validation.

NIST, AI Risk Management Framework, p. 11.

Archival Ethics

A code of ethics for archivists should establish high standards of conduct for the archival profession.

> It should introduce new members of the profession to those standards, remind experienced archivists of their professional responsibilities and inspire public confidence in the profession.

- **International Council on Archives**, Code of Ethics, “Introduction,” A.

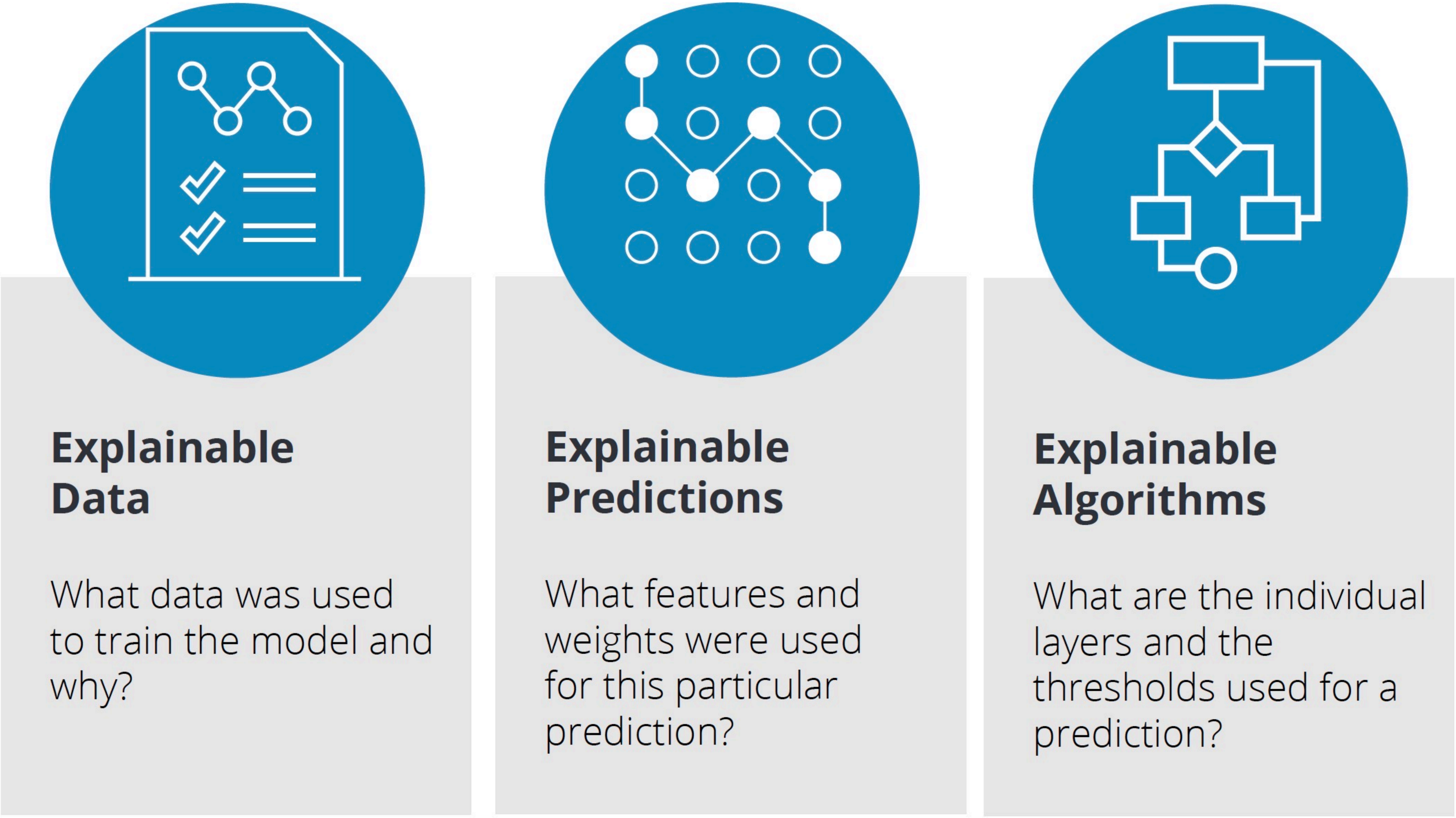
We exercise due caution and diligence in documenting and preserving the relationships between records and the activities that created them, as well as between records and the aggregations in which they belong, recognizing that these relationships are a necessary component of the records themselves.

We consider, analyze and evaluate the processes, methods, and technologies used to create, use and manage records with the intent of balancing our responsibility to optimize the value of records—and users’ access to them—against any risks and costs associated with doing so.

> We are mindful of, and document wherever possible, the biases inherent in records and information processing systems.

- **Association of Canadian Archivists**, Code of Ethics and Professional Conduct, and 1.a., 7, and 7.c.

Explainable AI



Explainable Data

What data was used to train the model and why?

Explainable Predictions

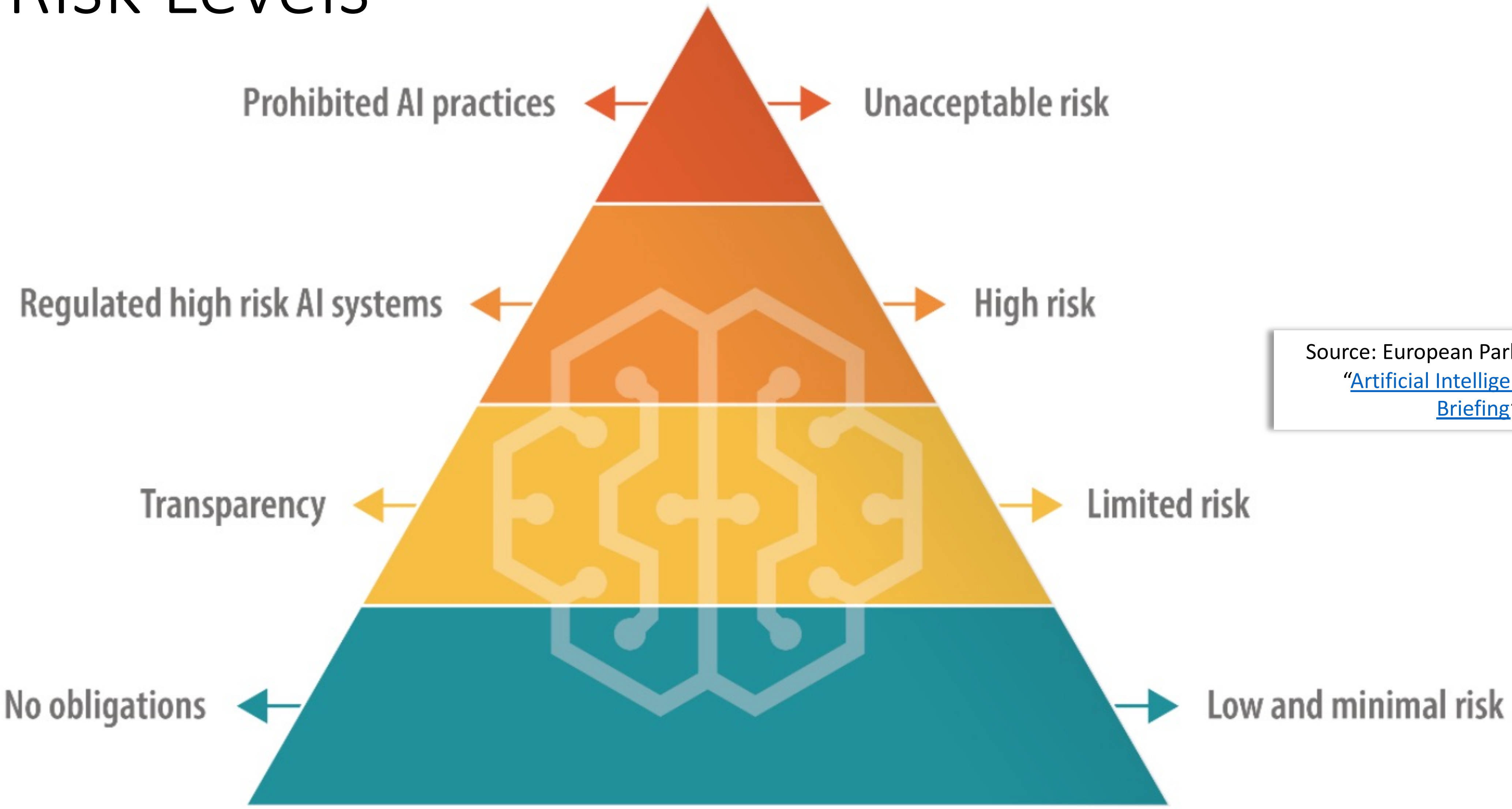
What features and weights were used for this particular prediction?

Explainable Algorithms

What are the individual layers and the thresholds used for a prediction?

Illustration from Amy E. Hodler, "AI & Graph Technology: AI Explainability" *Neo4j Blog* (26 August 2019).

AI Risk Levels



Source: European Parliament, "Artificial Intelligence Act: Briefing" (2022)

Records for AI Systems

[AI system developers and operators should document] the following information during the design, development and deployment phases, and retaining this documentation for a length of time appropriate to the decision type or industry:

- the provenance of the training data, the methods of collection and treatment, how the data was moved, and measures taken to maintain its accuracy over time;
- the model design and algorithms employed, and;
- changes to the codebase, and authorship of those changes

- Smart Dubai, AI Ethics Principles & Guidelines (2018), 1.3.1.2, p. 27

- [Describe] the way in which the training data was collected
- Institutions should use rigorous methods to validate their models and document those methods and results. In particular, they should routinely perform tests to assess and determine whether the model generates discriminatory harm. Institutions are encouraged to make the results of such tests public.

- ACM, Statement on Algorithmic Transparency and Accountability (2017), 5, 7, p. 2.

- "Archive of Systems Decisions" – a long-term source documenting a public authority's decisions on which systems do not need to prepare algorithmic impact assessments;
- A datasheet is a semi-structured document that asks questions like 'Why was the dataset created?,' 'How was the data collected?,' etc.

- European Parliamentary Research Services, A governance framework for algorithmic accountability and transparency (2019), pp. 54, 57.

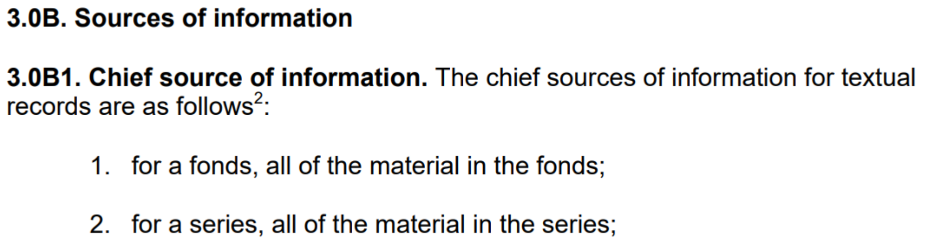
Researching Provenance, Descriptions

The fact is, we cannot take the records all together and know what went before and what followed after. We can, however, make the act of picking our text transparent to our users through description..”

- Heather MacNeil, “Picking Our Text: Archival Description, Authenticity, and the Archivist as Editor”, *American Archivist* #68 (2005), p. 278.

Recognizing that records originate in and are influenced by a complex interplay of legal, administrative, informational, and cultural factors over time, we strive to continuously improve our preservation and representation of these contexts.

- Association of Canadian Archivists, Code of Ethics and Professional Conduct (2017), 1.b.

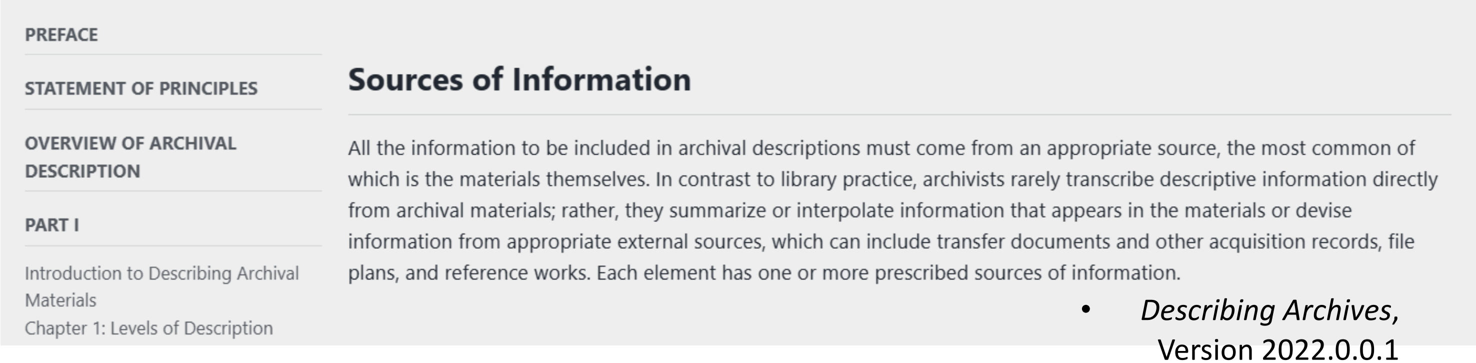


3.0B. Sources of information

3.0B1. Chief source of information. The chief sources of information for textual records are as follows[2]:

1. for a fonds, all of the material in the fonds;
2. for a series, all of the material in the series;

- *Rules for Archival Description,* Revised version (2008)



PREFACE
STATEMENT OF PRINCIPLES
OVERVIEW OF ARCHIVAL DESCRIPTION
PART I
Introduction to Describing Archival Materials
Chapter 1: Levels of Description

Sources of Information

All the information to be included in archival descriptions must come from an appropriate source, the most common of which is the materials themselves. In contrast to library practice, archivists rarely transcribe descriptive information directly from archival materials; rather, they summarize or interpolate information that appears in the materials or devise information from appropriate external sources, which can include transfer documents and other acquisition records, file plans, and reference works. Each element has one or more prescribed sources of information.

- *Describing Archives,* Version 2022.0.0.1

Responsible Use of AI Tools: PI detection

Type of PI	*Presidio* (Microsoft)	*Comprehend* (AWS)	*PII Tools*
Salary	x	x	x
Performance review	x	x	x
Religious views	✓	x	✓
Opinion	✓	x	x

- NB: *excerpts* of label sets. These tools can detect additional types of PI. Kudos to these vendors for making this information available!
- *Transcribe*, a subset of the *Comprehend* (AWS) label set, is described as "optimized for the financial sector".
- *Presidio* detection of religious views and opinions is dependent on "custom logic and context".

Perceptions of AI and Archival Identity

“The objectivity and impartiality of archivists is the measure of their professionalism.”

- International Council on Archives, Code of Ethics, #1

“We also acknowledge that archivists and archival practices are never neutral.”

- Society of American Archivists, Core Values Statement and Code of Ethics, “Overview”

“…despite the undeniably consistent picture that we see across studies of scientific data collection, the desire to remove the human from the data in order to enhance objectivity remains very strong. Invariably, it seems like the ethical move.”

- Melanie Feinberg, “The Myth of Objective Data” *The MIT Press Reader* (2022?)

Regulatory sandboxes?

Cool!!

Thank you!

Jim Suderman
InterPARES Trust AI Project

InterPARES
Trust AI