AI SUMMIT

COLUMBUS, OH • OCTOBER 25–27, 2022

BUILDING TRUST IN AI SYSTEMS: WHERE ARE WE NOW?

A DISCUSSION OF APPROACHES IN MOTION

AFDO RAPS HEALTHCARE PRODUCTS COLLABORATIVE

AFDO RAPS

Bakul Patel, Senior Director, Global Digital Health Strategy & Regulatory, Google

Koen Cobbaert, Senior Manager – Quality, Standards & Regulations, Philips

Rohit Nayak, Co-Founder CEO, Band Connect Inc. | CEO, Electronic Registry Systems, Inc.

Discussion Outline

- **Why Trust?**
  - Why is it important?
  - Different names and flavors - Transparency, Trustworthiness, Explainability
- **Driving Forces**
  - GDPR, XAI, EU AI ACT , AI Bill of Rights

This discussion reviews both the regulatory policy as well as the steps being taken by a significant industry player

AI SUMMIT
COLUMBUS, OH • OCTOBER 25–27, 2022

Driving Forces for Explainable AI

Update

AI Bill of Rights

Safe and Effective Systems

Algorithmic Discrimination Protections

Data Privacy

Notice and Explanation

Human Alternatives, Consideration, and Fallback

BLUEPRINT FOR AN
AI BILL OF RIGHTS
MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE

OCTOBER 2022

FROM PRINCIPLES TO PRACTICE

A TECHINCAL COMPANION TO THE BLUEPRINT FOR AN AI BILL OF RIGHTS

https://www.whitehouse.gov/ostp/ai-bill-of-rights/

Explainable AI

- EU GDPR – Right of Explanation
- IEEE - Standard for XAI – eXplainable Artificial Intelligence - for Achieving Clarity and Interoperability of AI Systems Design
- IEEE - Guide for an Architectural Framework for Explainable Artificial Intelligence
- DARPA’s XAI Initiative
- …..

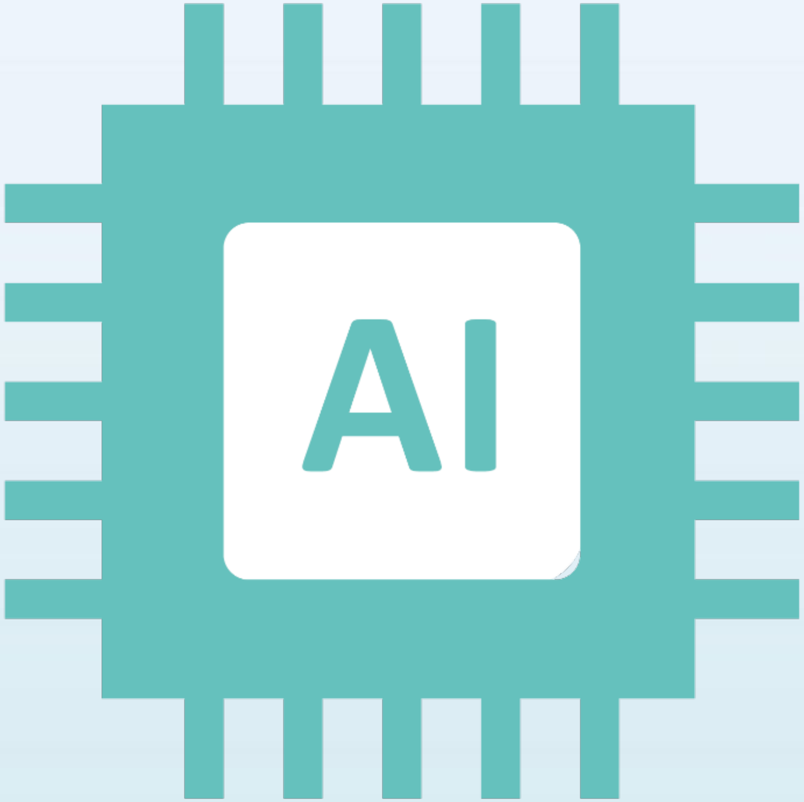
EU AI ACT

What is it?

Why are we talking about it?

Where does it stand?

legislative draft available
currently in legislative process
adoption expected Q2 2023
data of application Q2 2025 (?)



AI System – Commission Definition

Proposed AIA Art. 3(1)

an AI system is

software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.

Annex I (can be updated through delegated act)

a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;

b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

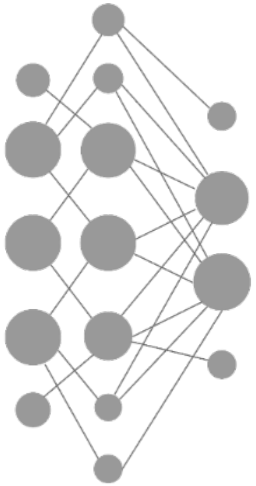
c) Statistical approaches, Bayesian estimation, search and optimization methods

definition

reads as

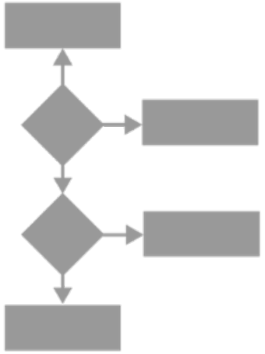
AI system = any software application

In listing technologies considered AI, Annex I tries to compensate for a vague AI system definition, but as technologies can be added or removed over time, it increases legal uncertainty



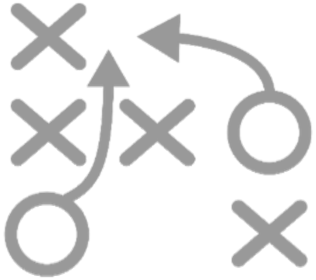
machine learning

inference and deductive engines



reasoning and expert systems

search & optimization methods



logic- and inductive programming

AI Act contains mandatory requirements for

High-Risk AI systems

=

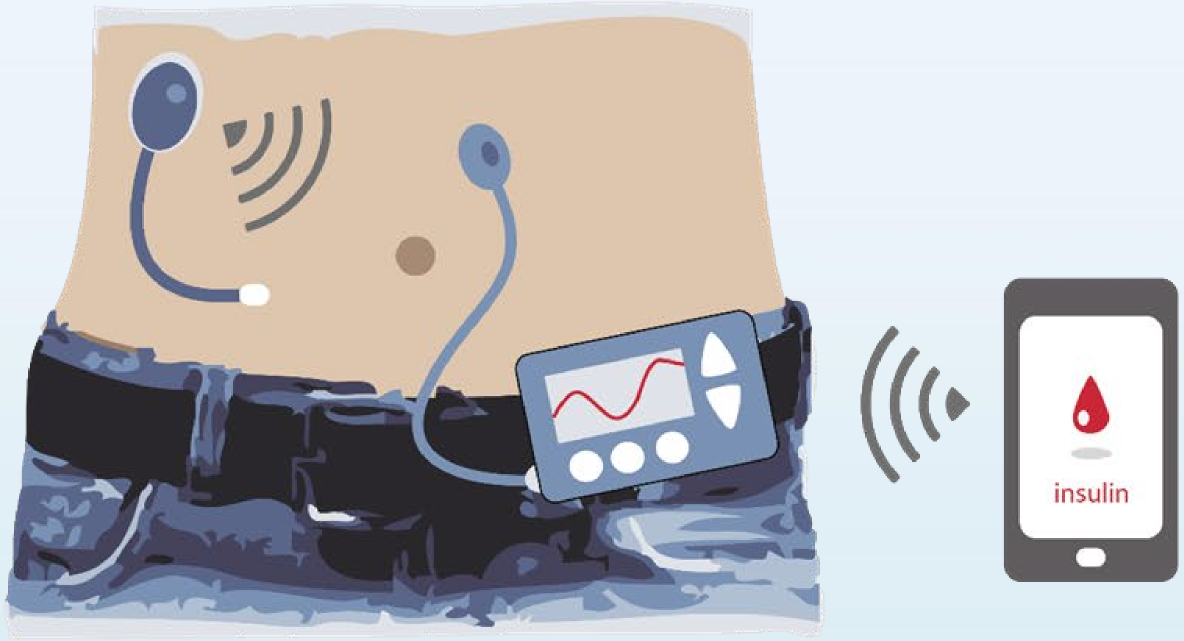
regulated products or safety components of regulated products
which are subject to third-party assessment under the relevant sectorial legislation

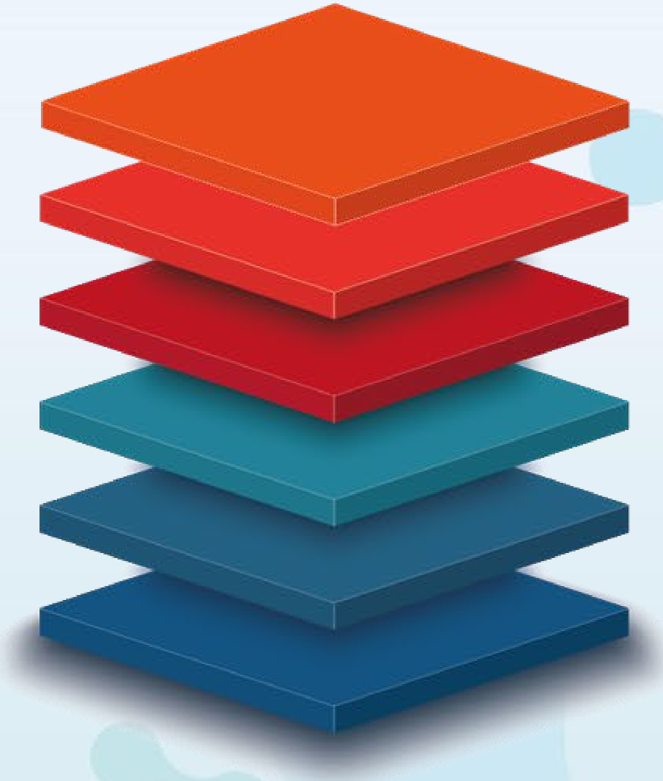
and for AI systems with transparency risks

Implication:

medical devices that are or that contain software as safety component and that are class IIa/B or higher are subject to AI Act

Legislative Lasagna





EU Data Act*

EU Health Data Space*

Artificial Intelligence Act*

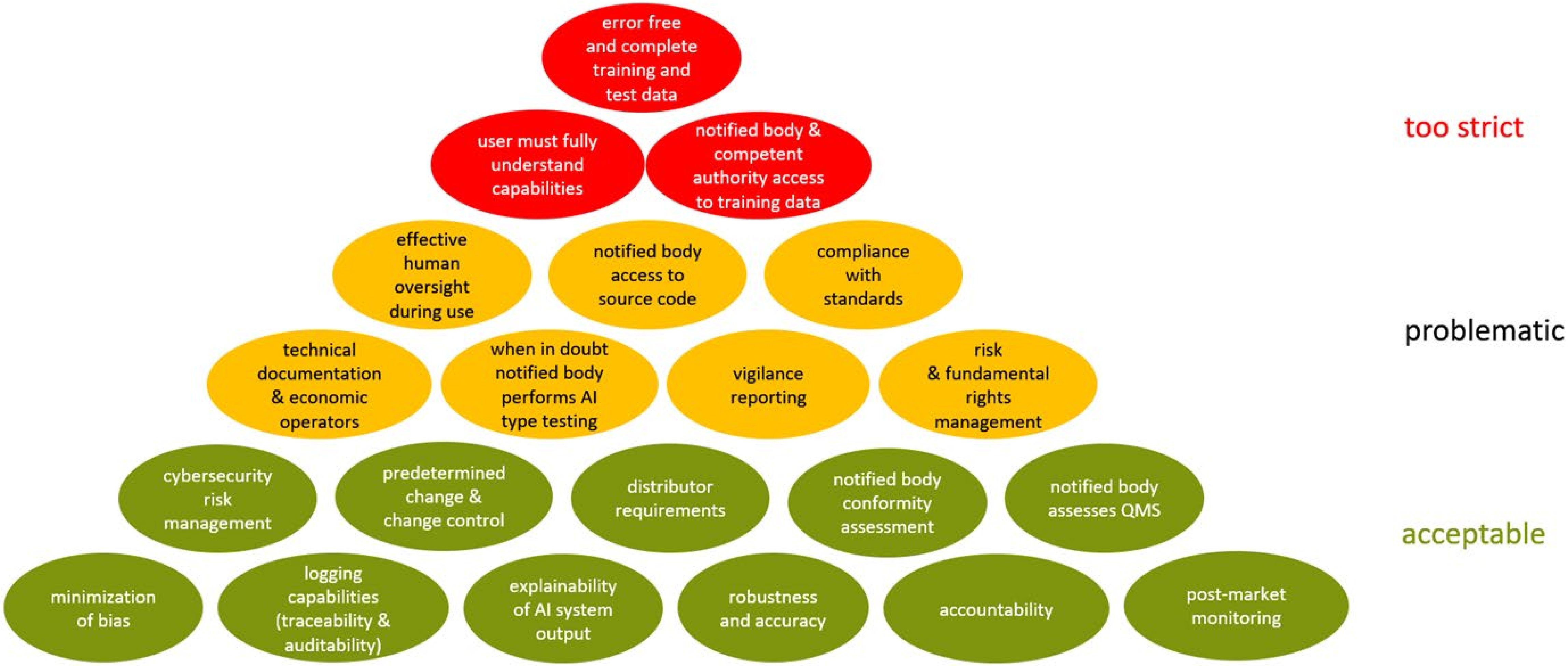
Radio Equipment Directive

Machinery Directive

Medical Device Regulation

Today, technical documentation of a closed-loop insulin pump needs to demonstrate compliance with three different legislations before the CE-mark can be assigned.

*In the future, three additional legislations may come on top.



error free and complete training and test data

user must fully understand capabilities

notified body & competent authority access to training data

too strict

effective human oversight during use

notified body access to source code

compliance with standards

technical documentation & economic operators

when in doubt notified body performs AI type testing

vigilance reporting

risk & fundamental rights management

problematic

cybersecurity risk management

predetermined change & change control

distributor requirements

notified body conformity assessment

notified body assesses QMS

acceptable

minimization of bias

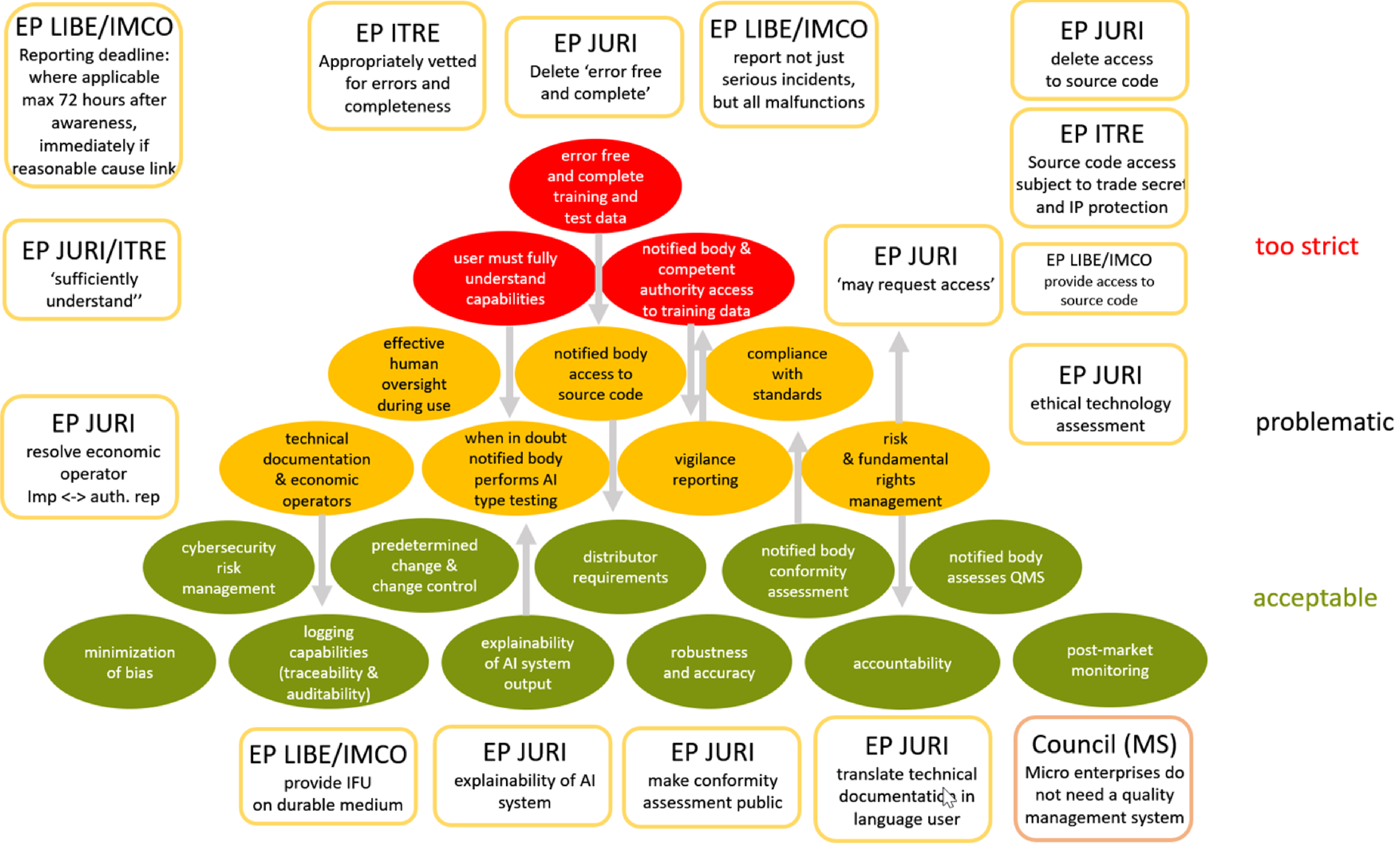
logging capabilities (traceability & auditability)

explainability of AI system output

robustness and accuracy

accountability

post-market monitoring



EP LIBE/IMCO
Reporting deadline: where applicable max 72 hours after awareness, immediately if reasonable cause link

EP ITRE
Appropriately vetted for errors and completeness

EP JURI
Delete 'error free and complete'

EP LIBE/IMCO
report not just serious incidents, but all malfunctions

EP JURI
delete access to source code

EP ITRE
Source code access subject to trade secret and IP protection

EP JURI/ITRE
'sufficiently understand''

EP JURI
'may request access'

EP LIBE/IMCO
provide access to source code

too strict

error free and complete training and test data

user must fully understand capabilities

notified body & competent authority access to training data

EP JURI
ethical technology assessment

problematic

effective human oversight during use

notified body access to source code

compliance with standards

EP JURI
resolve economic operator
Imp <-> auth. rep

technical documentation & economic operators

when in doubt notified body performs AI type testing

vigilance reporting

risk & fundamental rights management

acceptable

cybersecurity risk management

predetermined change & change control

distributor requirements

notified body conformity assessment

notified body assesses QMS

minimization of bias

logging capabilities (traceability & auditability)

explainability of AI system output

robustness and accuracy

accountability

post-market monitoring

EP LIBE/IMCO
provide IFU on durable medium

EP JURI
explainability of AI system

EP JURI
make conformity assessment public

EP JURI
translate technical documentation in language user

Council (MS)
Micro enterprises do not need a quality management system

Standards with high operationalization value

for implementing AI Act requirements

Overlaps & Conflicts: **extra costs, for little or no added value**

- **ISO/IEC 4213** Information technology — Artificial Intelligence — Assessment of ML classification performance
- **ISO/IEC 5259-3** Data quality for analytics and ML — Part 3: Data quality management requirements and guidelines
- **IEC 62304 <->** **ISO/IEC 5338** Information technology — Artificial intelligence — AI system life cycle processes
- **ISO/IEC 5469** Artificial intelligence — Functional safety and AI systems
- **ISO 14971 <->** **ISO/IEC 23894-2** Information Technology — Artificial Intelligence — Risk Management
- **ISO/IEC 24027** Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making
- **ISO IEC 24029-1** Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
- **ISO/IEC 38507** Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations
- **ISO 13485 <->** **ISO/IEC 42001** Information Technology — Artificial intelligence — Management system

List complied by AI Watch, joint initiative of European Commission and EC Joint Research Council
Above listed ISO/IEC SC42 standards are still under development

Lessons from Industry

Where are we headed?

Artificial Intelligence Principles @Google

1. Be socially beneficial.

2. Avoid creating or reinforcing unfair bias.

3. Be built and tested for safety.

4. Be accountable to people.

5. Incorporate privacy design principles.

6. Uphold high standards of scientific excellence.

7. Be made available for uses that accord with these principles.

ai.google/principles/

Which includes things we will not do

We will not pursue certain AI applications...

likely to cause overall harm

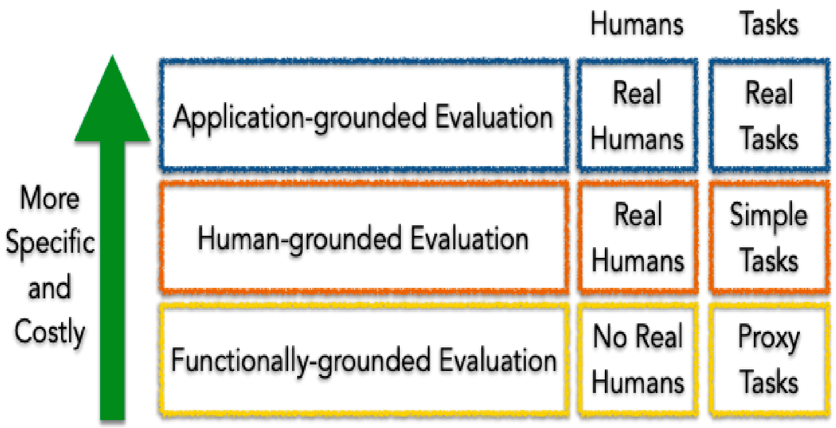
weapons or those that direct injury

surveillance violating internationally accepted norms

purpose contravenes international law and human rights

ai.google/principles/

AI SUMMIT
COLUMBUS, OH • OCTOBER 25–27, 2022



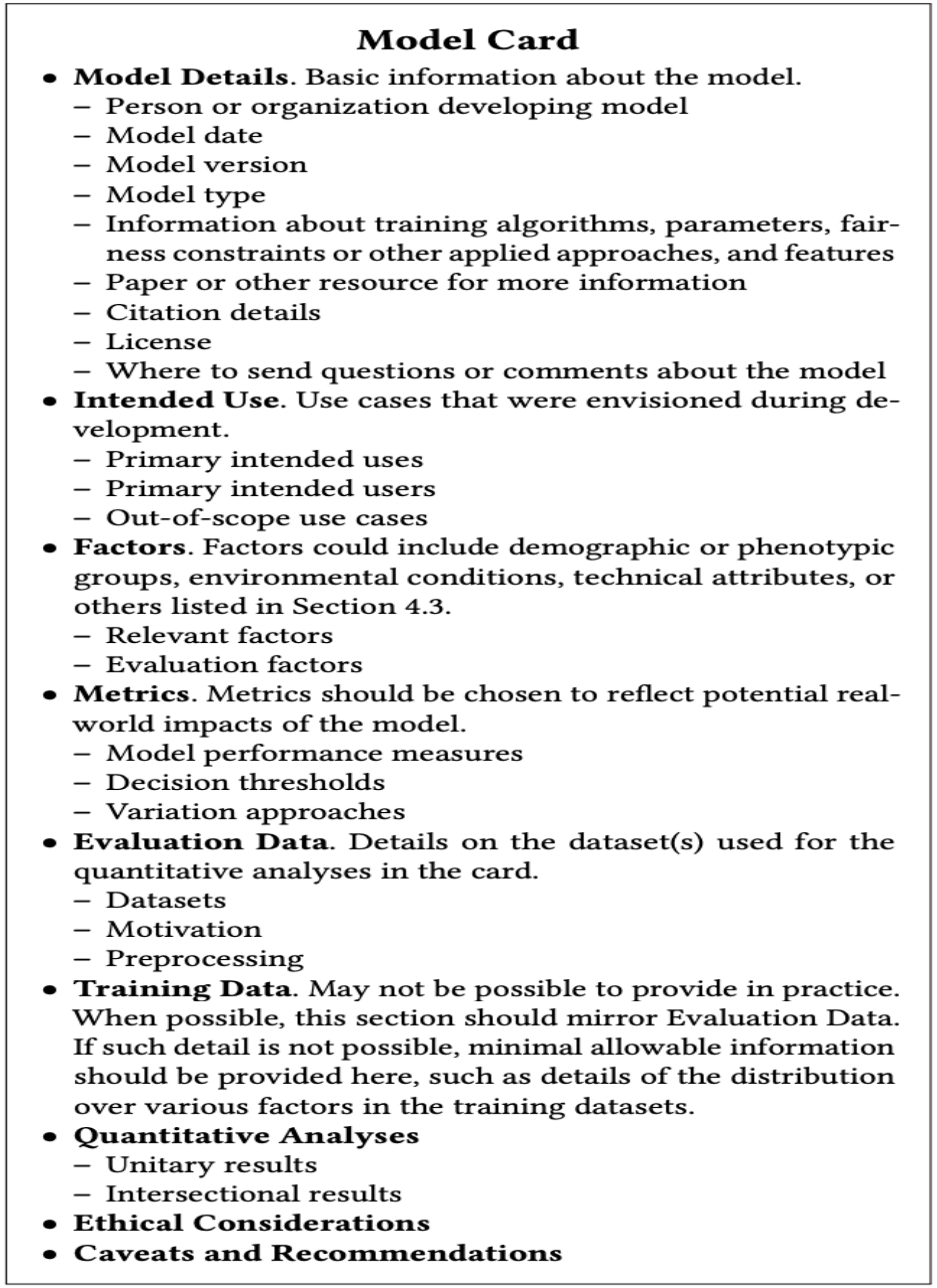
	Humans	Tasks
Application-grounded Evaluation	Real Humans	Real Tasks
Human-grounded Evaluation	Real Humans	Simple Tasks
Functionally-grounded Evaluation	No Real Humans	Proxy Tasks

Figure 1: Taxonomy of evaluation approaches for interpretability

https://arxiv.org/pdf/1702.08608.pdf

Model Cards for Model Reporting

https://arxiv.org/pdf/1810.03993.pdf



Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

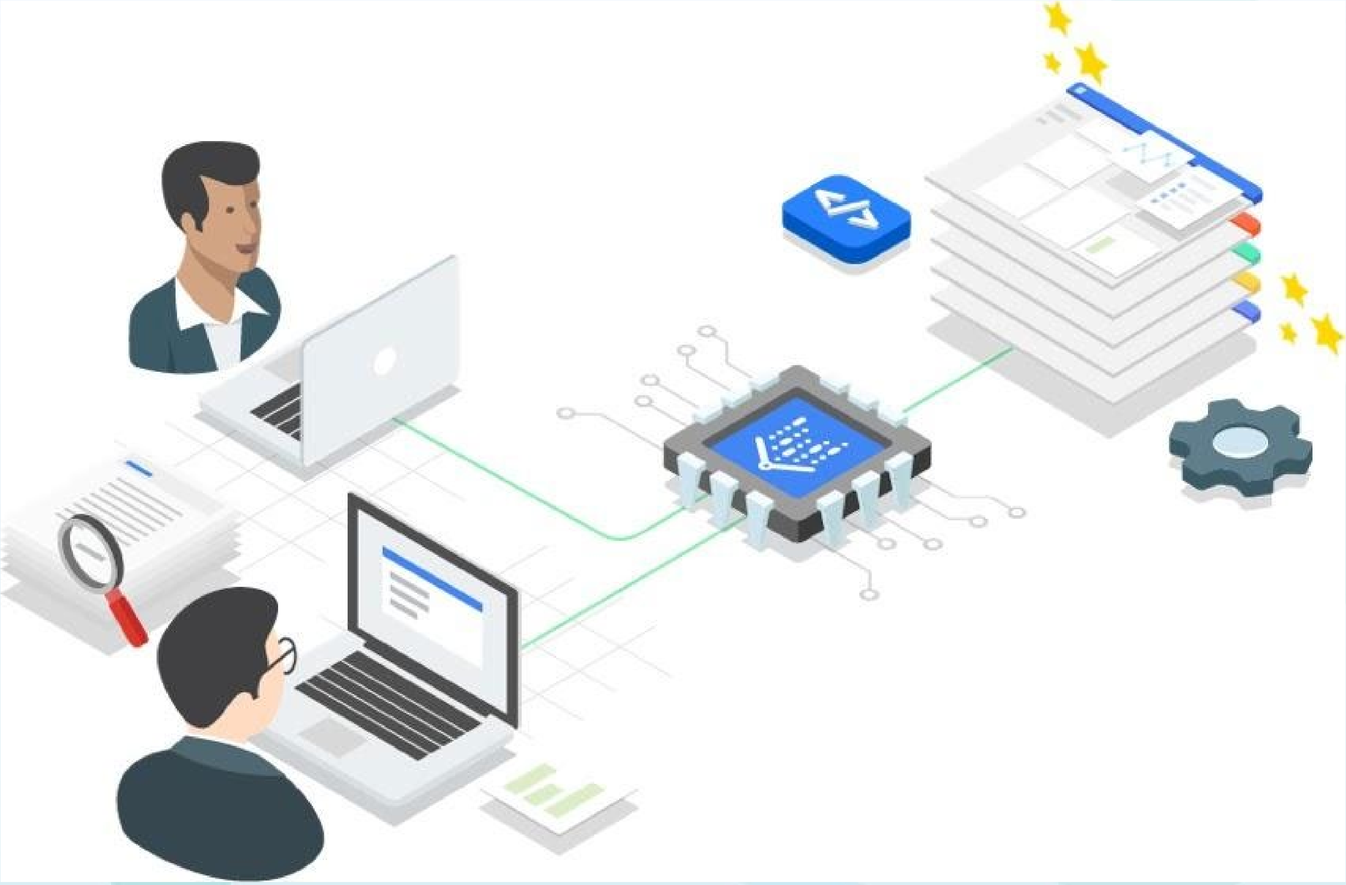
Figure 1: Summary of model card sections and suggested prompts for each.

Explainable AI

Understand AI output and build trust

Explainable AI is a set of tools and frameworks to help you understand and interpret predictions made by your machine learning models, natively integrated with a number of Google's products and services. With it, you can debug and improve model performance, and help others understand your models' behavior. You can also generate feature attributions for model predictions in AutoML Tables, BigQuery ML and Vertex AI, and visually investigate model behavior using the What-If Tool.

https://cloud.google.com/explainable-ai/



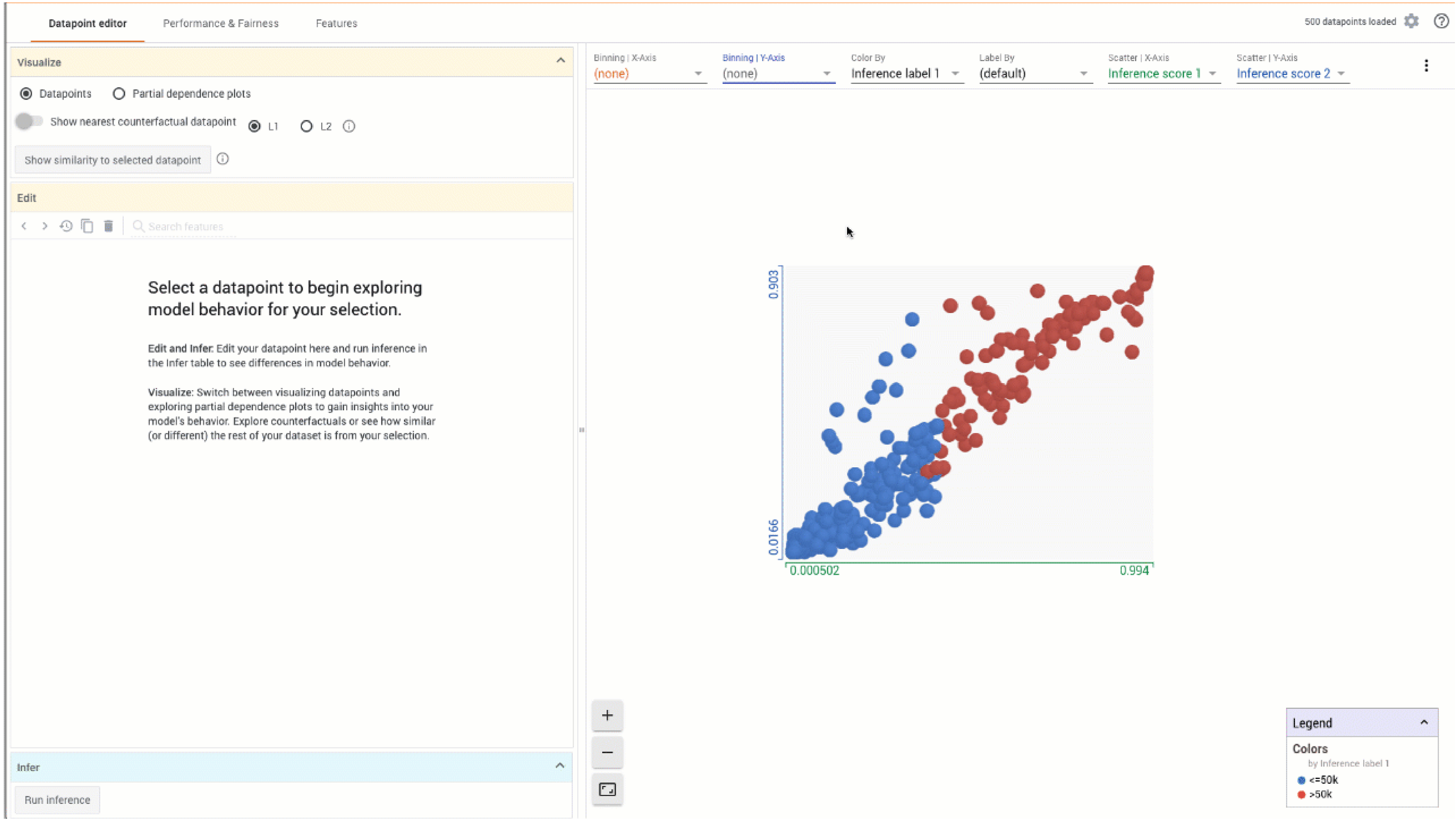
AI SUMM
COLUMBUS, OH • OCTOBER 25–

What - If Tool

Interactive, visual debugging of black - box models

[Website]

Probe classification and regression models, performing what - if analysis and analyzing fairness.



Select a datapoint to begin exploring model behavior for your selection.

Edit and Infer: Edit your datapoint here and run inference in the Infer table to see differences in model behavior.

Visualize: Switch between visualizing datapoints and exploring partial dependence plots to gain insights into your model's behavior. Explore counterfactuals or see how similar (or different) the rest of your dataset is from your selection.

500 datapoints loaded

Binning | X-Axis: (none)
Binning | Y-Axis: (none)
Color By: Inference label 1
Label By: (default)
Scatter | X-Axis: Inference score 1
Scatter | Y-Axis: Inference score 2

google_logo

AI SUMM

COLUMBUS, OH • OCTOBER 25–

Language Interpretability Tool

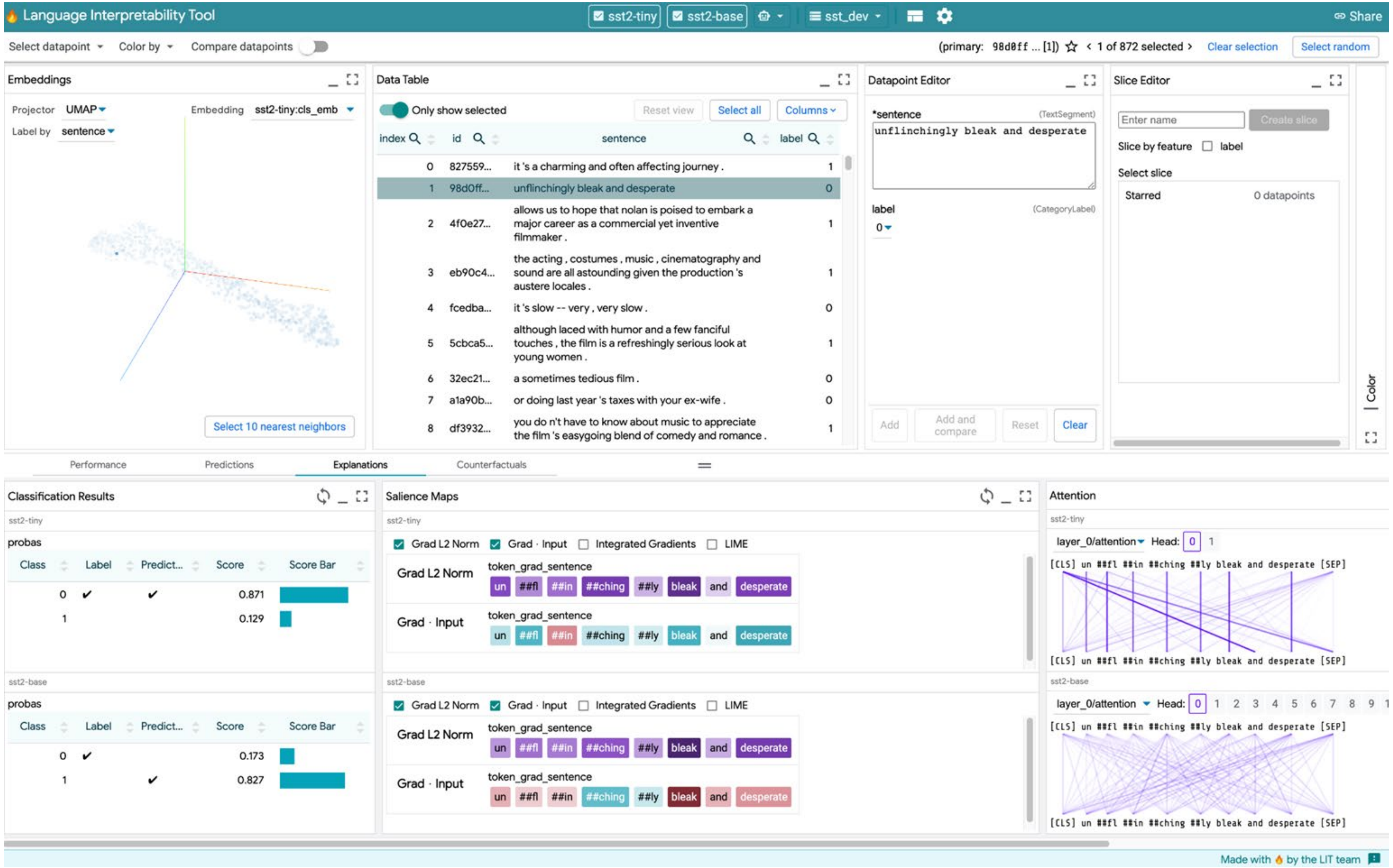
Interactive, extensible, visual debugging of NLP models and beyond

[Website]

Successor to the What - If Tool.

Probe models of all types (with a focus on NLP), explore model internals, prediction explanations, fairness, counterfactual generation, and more.

google_logo



index	id	sentence	label
0	827559...	it 's a charming and often affecting journey .	1
1	98d0ff...	unflinchingly bleak and desperate	0
2	4f0e27...	allows us to hope that nolan is poised to embark a major career as a commercial yet inventive filmmaker .	1
3	eb90c4...	the acting , costumes , music , cinematography and sound are all astounding given the production 's austere locales .	1
4	fcedba...	it 's slow -- very , very slow .	0
5	5cbca5...	although laced with humor and a few fanciful touches , the film is a refreshingly serious look at young women .	1
6	32ec21...	a sometimes tedious film .	0
7	a1a90b...	or doing last year 's taxes with your ex-wife .	0
8	df3932...	you do n't have to know about music to appreciate the film 's easygoing blend of comedy and romance .	1

*sentence (TextSegment)

unflinchingly bleak and desperate

label (CategoryLabel)

0

sst2-tiny

probas

Class	Label	Predict...	Score	Score Bar
0	✔	✔	0.871	
1			0.129	

sst2-base

probas

Class	Label	Predict...	Score	Score Bar
0	✔		0.173	
1		✔	0.827	

sst2-tiny

☑ Grad L2 Norm ☑ Grad · Input □ Integrated Gradients □ LIME

Grad L2 Norm — token_grad_sentence: un ##fl ##in ##ching ##ly bleak and desperate

Grad · Input — token_grad_sentence: un ##fl ##in ##ching ##ly bleak and desperate

sst2-base

☑ Grad L2 Norm ☑ Grad · Input □ Integrated Gradients □ LIME

Grad L2 Norm — token_grad_sentence: un ##fl ##in ##ching ##ly bleak and desperate

Grad · Input — token_grad_sentence: un ##fl ##in ##ching ##ly bleak and desperate

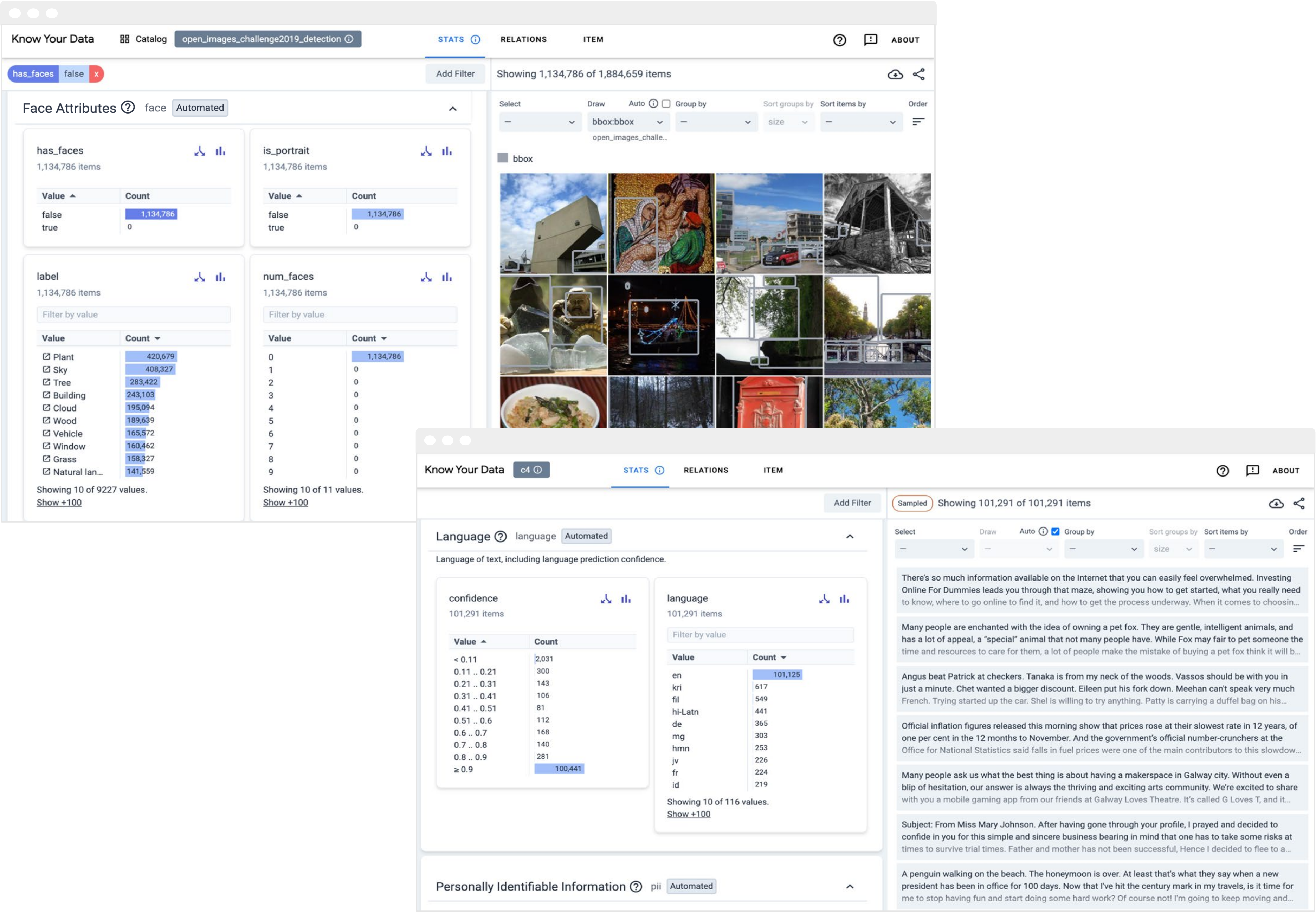
AI SUMM
COLUMBUS, OH • OCTOBER 25–

Know Your Data

KYD is an ML-based dataset exploration tools for rich, unstructured data

- Automatically computes signals
- Surface most biased data feature automatically through sorting and coloring

google_logo



Know Your Data | Catalog | open_images_challenge2019_detection | STATS | RELATIONS | ITEM | ABOUT

has_faces false x | Add Filter | Showing 1,134,786 of 1,884,659 items

Select — | Draw bbox:bbox open_images_challe... | Auto ☐ | Group by — | Sort groups by size | Sort items by — | Order

bbox

Face Attributes face Automated

has_faces
1,134,786 items

Value	Count
false	1,134,786
true	0

is_portrait
1,134,786 items

Value	Count
false	1,134,786
true	0

label
1,134,786 items

Filter by value

Value	Count
Plant	420,679
Sky	408,327
Tree	283,422
Building	243,103
Cloud	195,094
Wood	189,639
Vehicle	165,572
Window	160,462
Grass	158,327
Natural lan...	141,559

Showing 10 of 9227 values.
Show +100

num_faces
1,134,786 items

Filter by value

Value	Count
0	1,134,786
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0

Showing 10 of 11 values.
Show +100

Know Your Data | c4 | STATS | RELATIONS | ITEM | ABOUT

Add Filter | Sampled | Showing 101,291 of 101,291 items

Select — | Draw — | Auto ☑ | Group by — | Sort groups by size | Sort items by — | Order

Language language Automated

Language of text, including language prediction confidence.

confidence
101,291 items

Value	Count
< 0.11	2,031
0.11 .. 0.21	300
0.21 .. 0.31	143
0.31 .. 0.41	106
0.41 .. 0.51	81
0.51 .. 0.6	112
0.6 .. 0.7	168
0.7 .. 0.8	140
0.8 .. 0.9	281
≥ 0.9	100,441

language
101,291 items

Filter by value

Value	Count
en	101,125
kri	617
fil	549
hi-Latn	441
de	365
mg	303
hmn	253
jv	226
fr	224
id	219

Showing 10 of 116 values.
Show +100

Personally Identifiable Information pii Automated

There's so much information available on the Internet that you can easily feel overwhelmed. Investing Online For Dummies leads you through that maze, showing you how to get started, what you really need to know, where to go online to find it, and how to get the process underway. When it comes to choosin...

Many people are enchanted with the idea of owning a pet fox. They are gentle, intelligent animals, and has a lot of appeal, a "special" animal that not many people have. While Fox may fair to pet someone the time and resources to care for them, a lot of people make the mistake of buying a pet fox think it will b...

Angus beat Patrick at checkers. Tanaka is from my neck of the woods. Vassos should be with you in just a minute. Chet wanted a bigger discount. Eileen put his fork down. Meehan can't speak very much French. Trying started up the car. Shel is willing to try anything. Patty is carrying a duffel bag on his...

Official inflation figures released this morning show that prices rose at their slowest rate in 12 years, of one per cent in the 12 months to November. And the government's official number-crunchers at the Office for National Statistics said falls in fuel prices were one of the main contributors to this slowdow...

Many people ask us what the best thing is about having a makerspace in Galway city. Without even a blip of hesitation, our answer is always the thriving and exciting arts community. We're excited to share with you a mobile gaming app from our friends at Galway Loves Theatre. It's called G Loves T, and it...

Subject: From Miss Mary Johnson. After having gone through your profile, I prayed and decided to confide in you for this simple and sincere business bearing in mind that one has to take some risks at times to survive trial times. Father and mother has not been successful, Hence I decided to flee to a...

A penguin walking on the beach. The honeymoon is over. At least that's what they say when a new president has been in office for 100 days. Now that I've hit the century mark in my travels, is it time for me to stop having fun and start doing some hard work? Of course not! I'm going to keep moving and...

Fairness Indicators

Open-source library that enables users to evaluate model performance for specific user groups ("sliced" analysis):

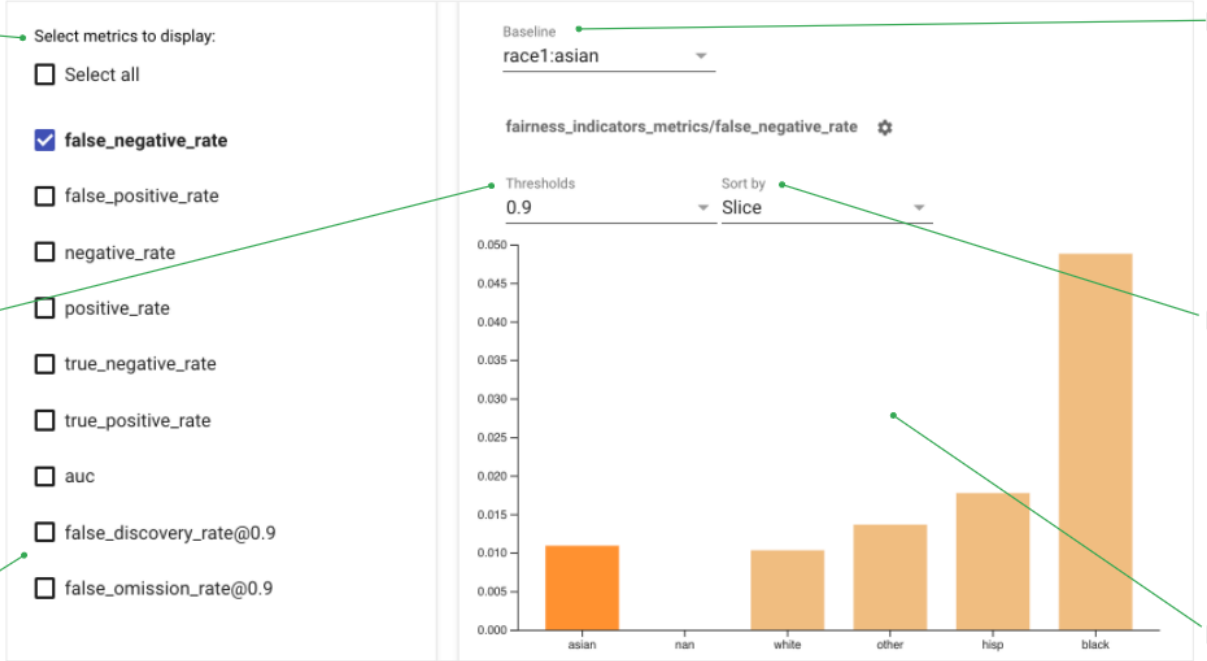
- Comes pre-loaded with common fairness metrics
- Provides interactive dashboard for rapid analysis & sharing insights with others
- Run analyses and visualize results in Jupyter notebooks or as part of TFX pipelines

Fairness Indicators dashboard

Select from common fairness metrics one or multiple

Set different decisions threshold(s)

Hover over metrics to inspect their definition



Select metrics to display:

☐ Select all

☑ **false_negative_rate**

☐ false_positive_rate

☐ negative_rate

☐ positive_rate

☐ true_negative_rate

☐ true_positive_rate

☐ auc

☐ false_discovery_rate@0.9

☐ false_omission_rate@0.9

Baseline: race1:asian

fairness_indicators_metrics/false_negative_rate

Thresholds: 0.9

Sort by: Slice

Select a baseline to compare performance against

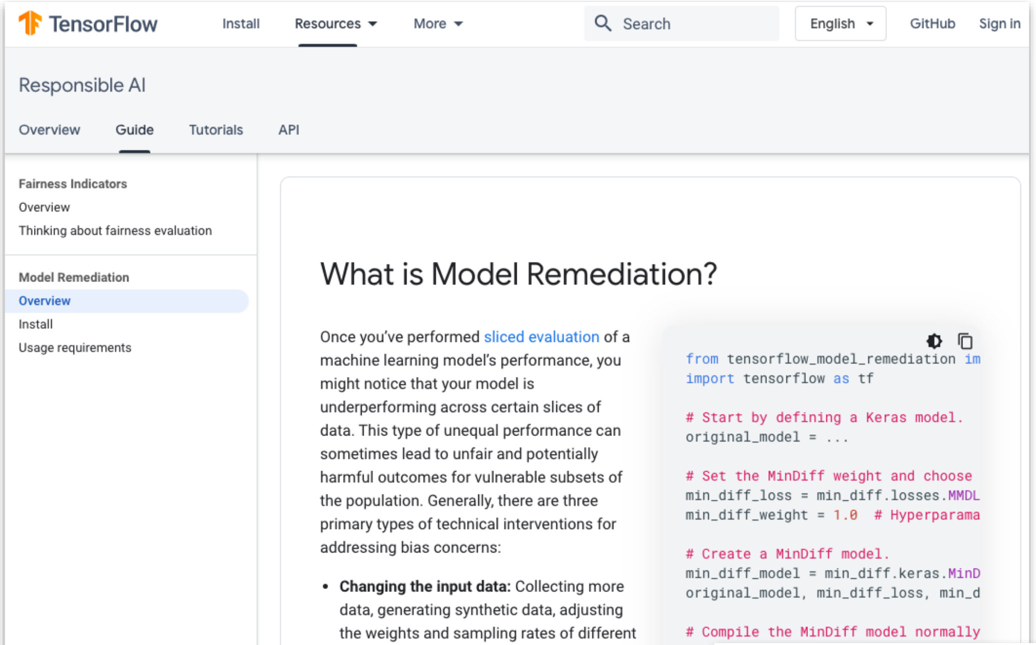
Choose slice(s) to display (e.g. subgroups)

View results for your selections, side-by-side

Model Remediation Library

[Open-source library](#) that enables users to train classifiers that equalize performance (provide "equal treatment") across a dimension, e.g. demographic group

Based on MinDiff modeling method (paper: *[Toward a better trade-off between performance and fairness with kernel-based distribution matching](#)*)

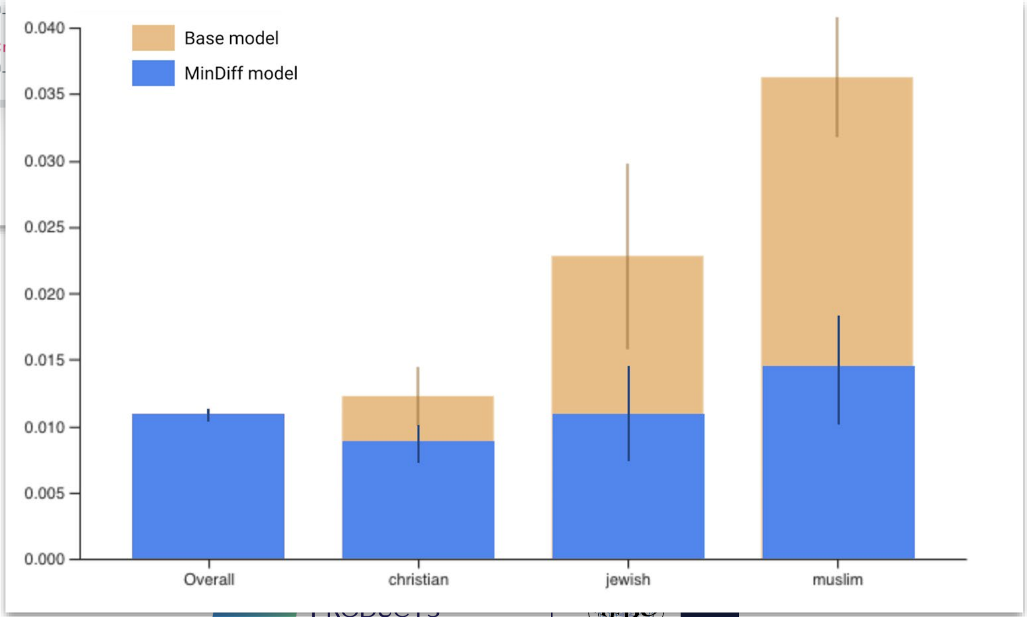


What is Model Remediation?

Once you've performed sliced evaluation of a machine learning model's performance, you might notice that your model is underperforming across certain slices of data. This type of unequal performance can sometimes lead to unfair and potentially harmful outcomes for vulnerable subsets of the population. Generally, there are three primary types of technical interventions for addressing bias concerns:

- **Changing the input data:** Collecting more data, generating synthetic data, adjusting the weights and sampling rates of different slices, etc.[1]
- **Intervening on the model:** Changing the model itself by introducing or altering model objectives, adding constraints, etc.[2]
- **Post-processing the results:** Modifying the outputs of the model or the interpretation of the outputs to improve performance across metrics.[3]

```
from tensorflow_model_remediation im
import tensorflow as tf

# Start by defining a Keras model.
original_model = ...

# Set the MinDiff weight and choose
min_diff_loss = min_diff.losses.MMDL
min_diff_weight = 1.0  # Hyperparama

# Create a MinDiff model.
min_diff_model = min_diff.keras.MinD
original_model, min_diff_loss, min_d

# Compile the MinDiff model normally
min_

# Cr
min_
```

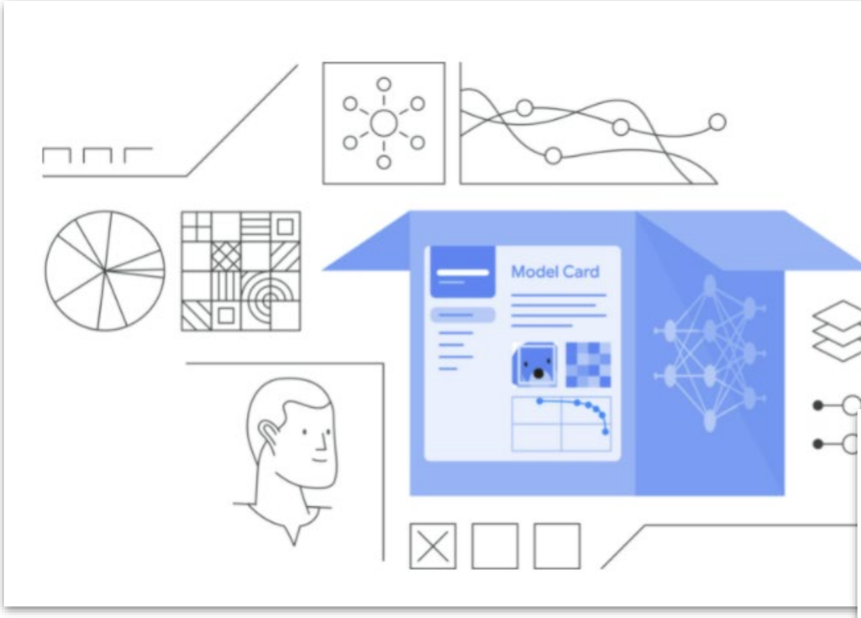


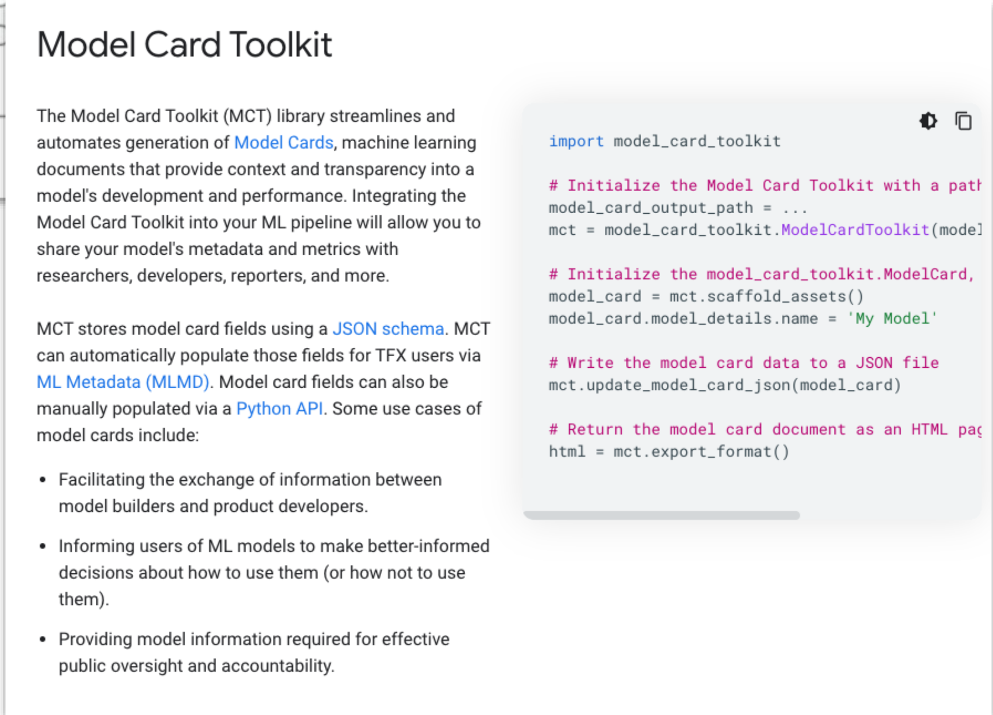
Model Cards

Model Cards offer a transparency framework for organizing & communicating key information about a model in a standardized way.

Open-source Model Card Toolkit library facilitates and streamlines the creation of model cards

Model Cards for Model Reporting paper (2019)





Model Card Toolkit

The Model Card Toolkit (MCT) library streamlines and automates generation of Model Cards, machine learning documents that provide context and transparency into a model's development and performance. Integrating the Model Card Toolkit into your ML pipeline will allow you to share your model's metadata and metrics with researchers, developers, reporters, and more.

MCT stores model card fields using a JSON schema. MCT can automatically populate those fields for TFX users via ML Metadata (MLMD). Model card fields can also be manually populated via a Python API. Some use cases of model cards include:

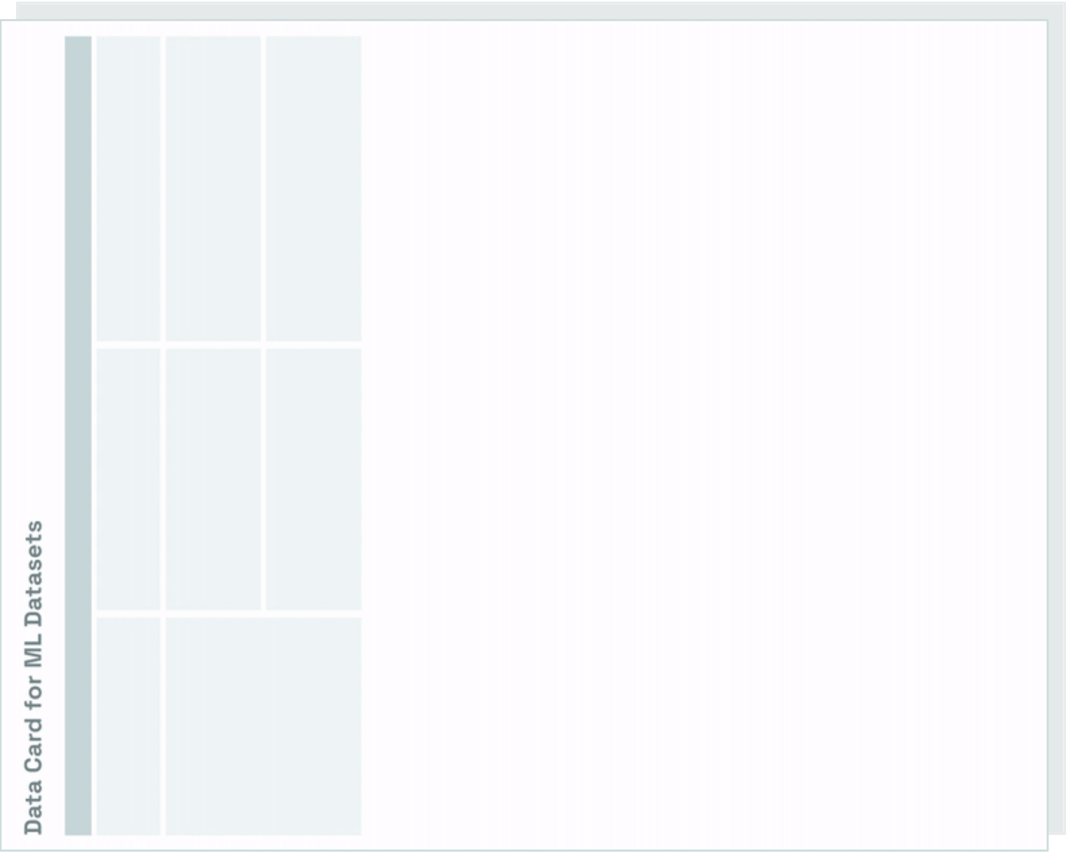
- Facilitating the exchange of information between model builders and product developers.
- Informing users of ML models to make better-informed decisions about how to use them (or how not to use them).
- Providing model information required for effective public oversight and accountability.

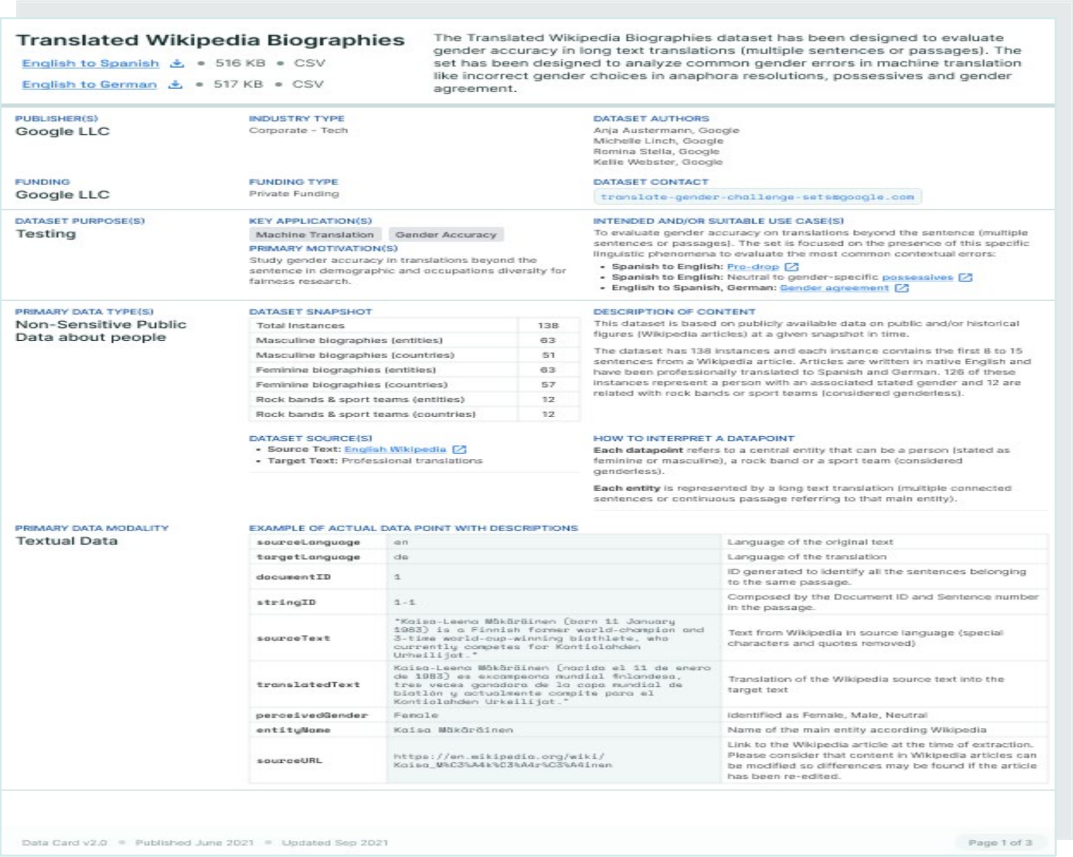
```
import model_card_toolkit

# Initialize the Model Card Toolkit with a path
model_card_output_path = ...
mct = model_card_toolkit.ModelCardToolkit(mode

# Initialize the model_card_toolkit.ModelCard,
model_card = mct.scaffold_assets()
model_card.model_details.name = 'My Model'

# Write the model card data to a JSON file
mct.update_model_card_json(model_card)

# Return the model card document as an HTML pag
html = mct.export_format()
```

Data Cards

Data Cards offer a structured way to document datasets & facilitate informed decision making for various stakeholders.

The Data Cards Playbook is a people-centered resource to help teams create customizable dataset documentation.





Translated Wikipedia Biographies

English to Spanish • 516 KB • CSV

English to German • 517 KB • CSV

The Translated Wikipedia Biographies dataset has been designed to evaluate gender accuracy in long text translations (multiple sentences or passages). The set has been designed to analyze common gender errors in machine translation like incorrect gender choices in anaphora resolutions, possessives and gender agreement.

PUBLISHER(S)
Google LLC

INDUSTRY TYPE
Corporate - Tech

DATASET AUTHORS
Anja Austermann, Google
Michelle Linch, Google
Romina Stella, Google
Kellie Webster, Google

FUNDING
Google LLC

FUNDING TYPE
Private Funding

DATASET CONTACT
translate-gender-challenge-sets@google.com

DATASET PURPOSE(S)
Testing

KEY APPLICATION(S)
Machine Translation Gender Accuracy

PRIMARY MOTIVATION(S)
Study gender accuracy in translations beyond the sentence in demographic and occupations diversity for fairness research.

INTENDED AND/OR SUITABLE USE CASE(S)
To evaluate gender accuracy on translations beyond the sentence (multiple sentences or passages). The set is focused on the presence of this specific linguistic phenomena to evaluate the most common contextual errors:

- **Spanish to English:** Pro-drop
- **Spanish to English:** Neutral to gender-specific possessives
- **English to Spanish, German:** Gender agreement

PRIMARY DATA TYPE(S)
Non-Sensitive Public Data about people

DATASET SNAPSHOT

Total Instances	138
Masculine biographies (entities)	63
Masculine biographies (countries)	51
Feminine biographies (entities)	63
Feminine biographies (countries)	57
Rock bands & sport teams (entities)	12
Rock bands & sport teams (countries)	12

DESCRIPTION OF CONTENT
This dataset is based on publicly available data on public and/or historical figures (Wikipedia articles) at a given snapshot in time.

The dataset has 138 instances and each instance contains the first 8 to 15 sentences from a Wikipedia article. Articles are written in native English and have been professionally translated to Spanish and German. 126 of these instances represent a person with an associated stated gender and 12 are related with rock bands or sport teams (considered genderless).

DATASET SOURCE(S)

- **Source Text:** English Wikipedia
- **Target Text:** Professional translations

HOW TO INTERPRET A DATAPOINT
Each datapoint refers to a central entity that can be a person (stated as feminine or masculine), a rock band or a sport team (considered genderless).

Each entity is represented by a long text translation (multiple connected sentences or continuous passage referring to that main entity).

PRIMARY DATA MODALITY
Textual Data

EXAMPLE OF ACTUAL DATA POINT WITH DESCRIPTIONS

sourceLanguage	en	Language of the original text
targetLanguage	de	Language of the translation
documentID	1	ID generated to identify all the sentences belonging to the same passage.
stringID	1-1	Composed by the Document ID and Sentence number in the passage.
sourceText	"Kaisa-Leena Mäkäräinen (born 11 January 1983) is a Finnish former world-champion and 3-time world-cup-winning biathlete, who currently competes for Kontiolahden Urheilijat."	Text from Wikipedia in source language (special characters and quotes removed)
translatedText	Kaisa-Leena Mäkäräinen (nacida el 11 de enero de 1983) es excampeona mundial finlandesa, tres veces ganadora de la copa mundial de biatlón y actualmente compite para el Kontiolahden Urheilijat."	Translation of the Wikipedia source text into the target text
perceivedGender	Female	Identified as Female, Male, Neutral
entityName	Kaisa Mäkäräinen	Name of the main entity according Wikipedia
sourceURL	https://en.wikipedia.org/wiki/Kaisa_M%C3%A4k%C3%A4r%C3%A4inen	Link to the Wikipedia article at the time of extraction. Please consider that content in Wikipedia articles can be modified so differences may be found if the article has been re-edited.

Data Card v2.0 • Published June 2021 • Updated Sep 2021

Page 1 of 3

AI SUMMIT
COLUMBUS, OH • OCTOBER 25–27, 2022

DISCUSSION

BUILDING TRUST IN AI SYSTEMS: WHERE ARE WE NOW?