



# OECD & TRUSTWORTHY AI: From Principles to Practice






*6 December 2022*  
*Lisbon Council*  
*Karine Perset*

Ensuring that AI  
safeguards are  
interoperable  
internationally is  
urgent

- Over- and under-regulation would be problematic
- AI is global and impacts everyone
- Regulations and standards need to be interoperable
- The window of opportunity to seek the right regulatory balance and ensure interoperability is now

# The OECD AI Principles

## 5 values-based principles for trustworthy, human-centric AI

-  Benefit People & Planet
-  Human rights, values & fairness
-  Transparent & explainable
-  Robust, secure & safe
-  Accountable

## 5 principles for national policies, for AI ecosystems to benefit societies

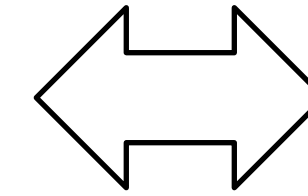
-  AI research & development
-  Data, compute, technologies
-  Policy & regulatory environment
-  Jobs & skills, labour transitions
-  International cooperation & measurement

# Putting the OECD AI Principles into practice

OECD Committee on Digital Economy Policy (CDEP)

## OECD Working Party on AI Governance (AIGO)

*National AI policies, regulatory sandboxes, jobs & skills, foresight etc.*



Other OECD committees and bodies

## Multi-stakeholder OECD.AI Network of Experts on AI (ONE AI)

*Expert group on AI risk & accountability*

*Expert group on AI incident reporting & tracking*

*Expert group on AI compute & climate*

*Expert group on AI foresight*

## OECD.AI Policy Observatory

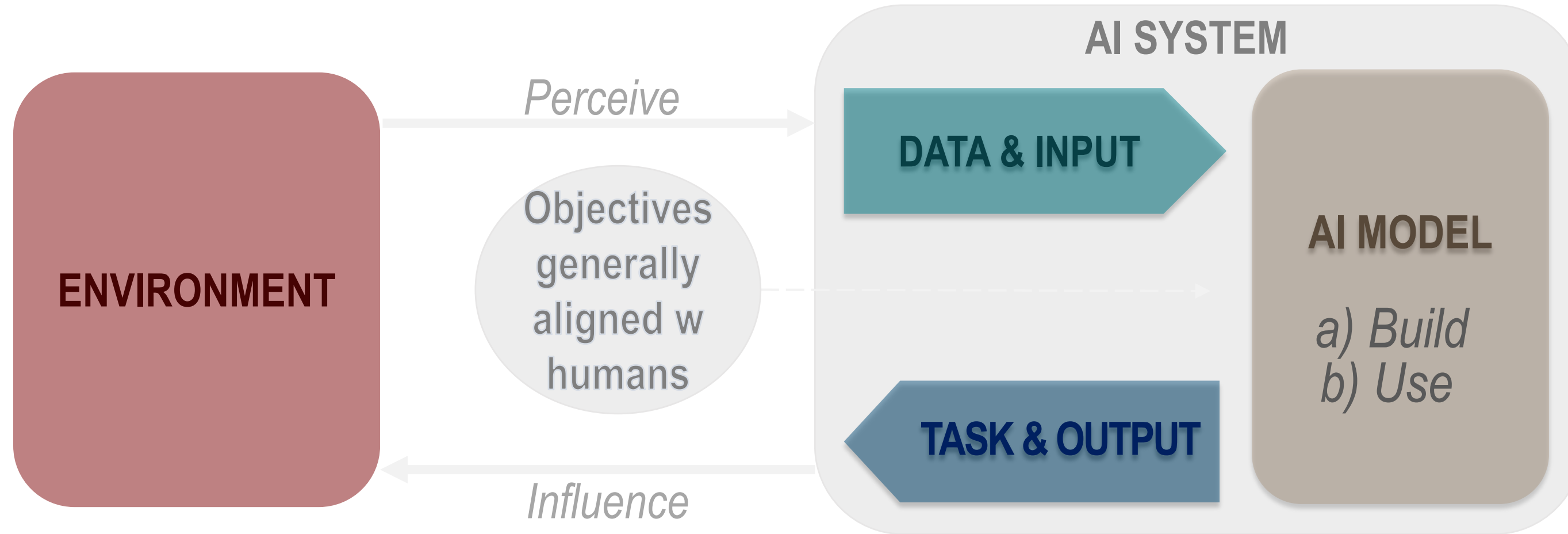
*A platform to share & shape public policies for trustworthy AI*

- database of national AI policies,*
- trends and data*
- repository of tools for trustworthy AI, etc.*





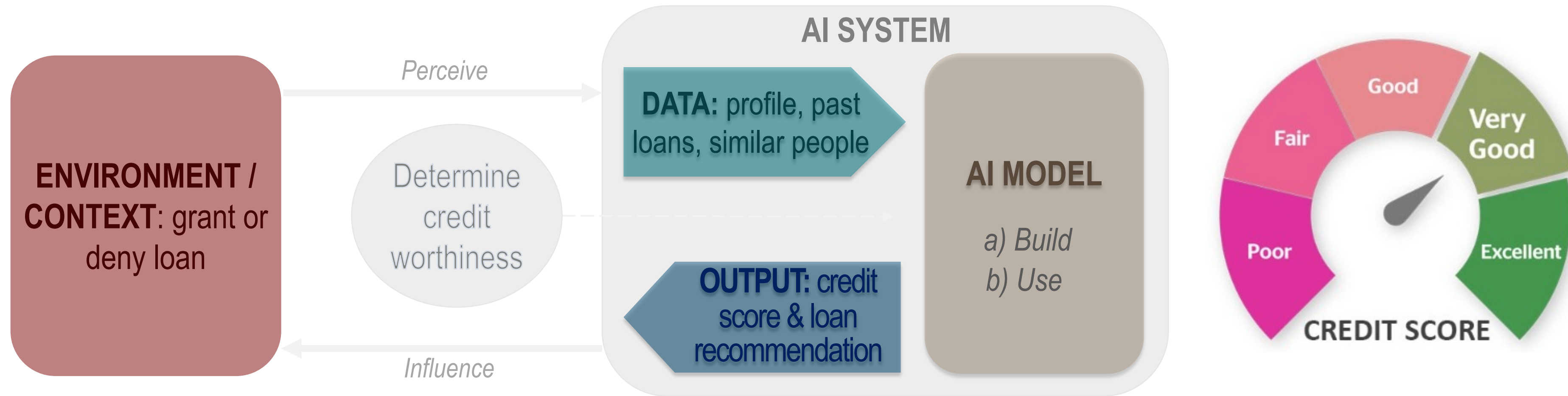
**WHAT IS AI?**... An AI system is a **machine-based** system that can...



for given **objectives** generally aligned with human preferences:

- perceive the environment using **data & inputs**;
- abstract these perceptions to **build a model** of the environment;
- **use the model** to generate **outputs** or to conduct **tasks**, such as predictions, recommendations or interactions;
- that **influence the environment** with more or less autonomy.

For example, a credit scoring AI system can...



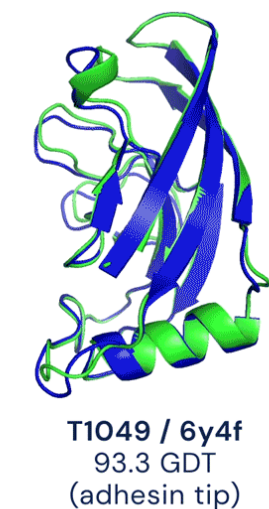
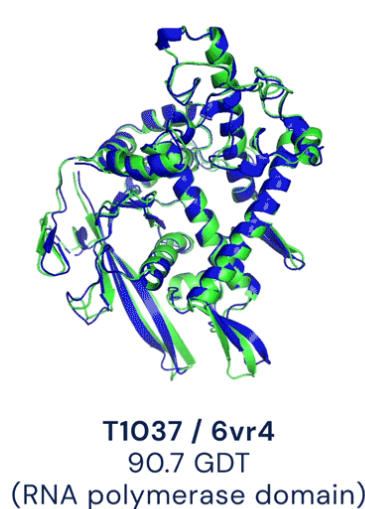
for a human-defined **objective** of determining credit-worthiness:

- perceive context using **data** on people, their past loans & similar people;
- abstract these perceptions to **build** a statistical credit-scoring model;
- **use the model** to generate a credit score and loan recommendation output
- that **influences** whether someone is granted or denied a loan with more or less autonomy / involvement of human bankers.

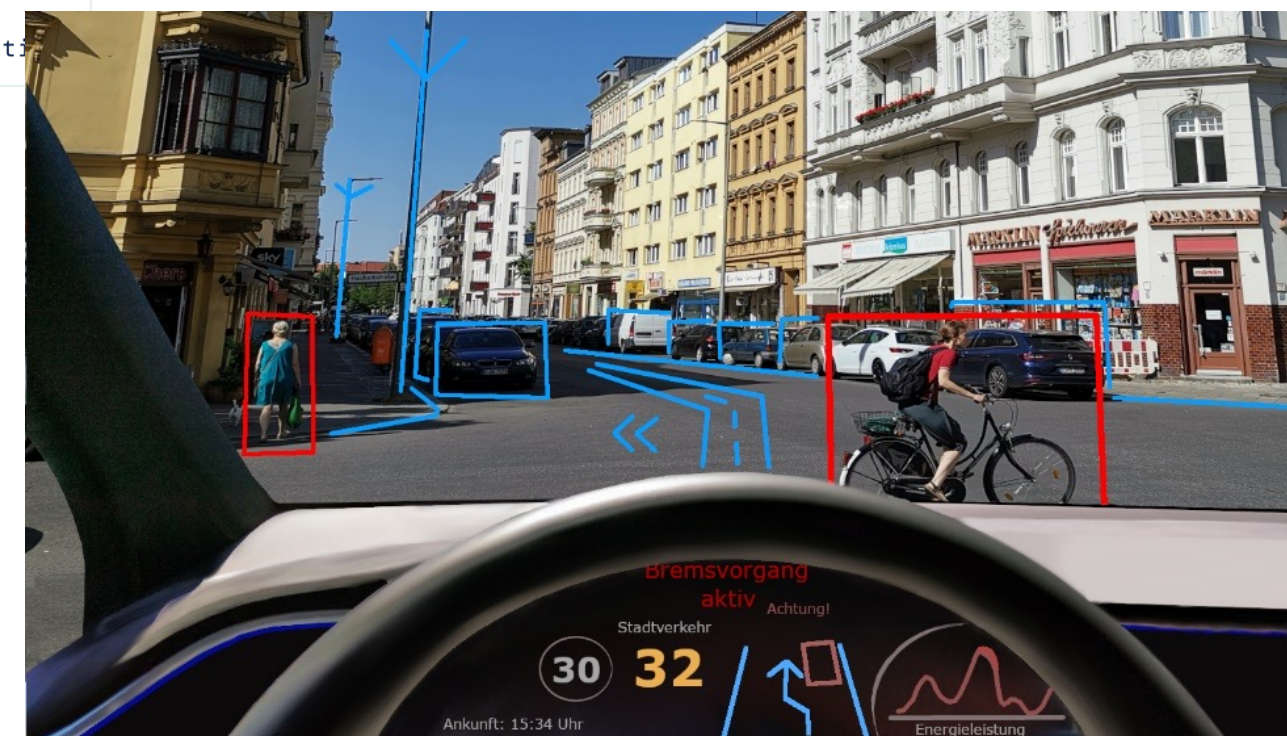
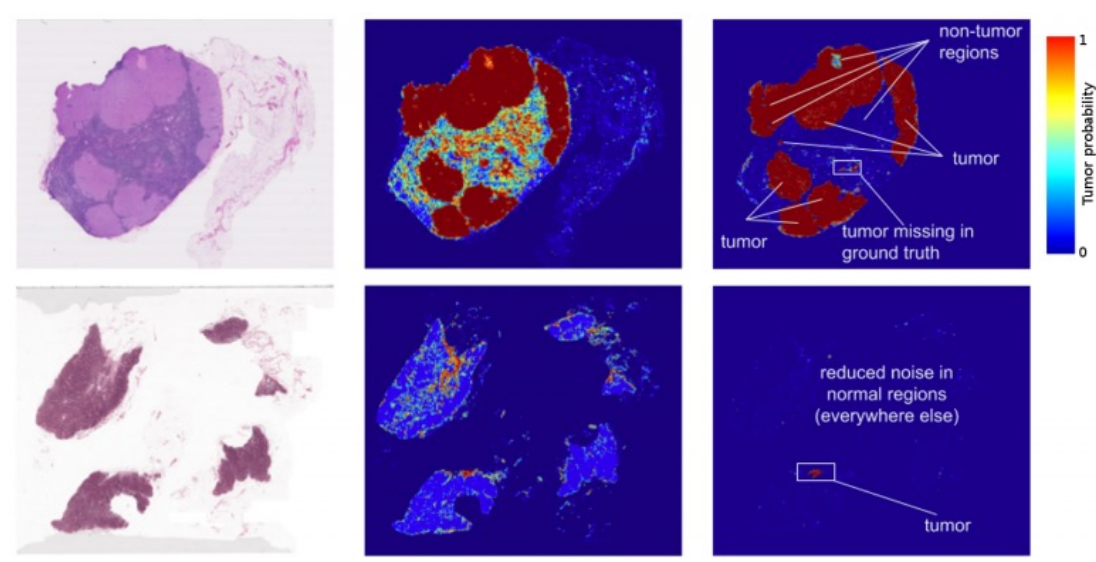
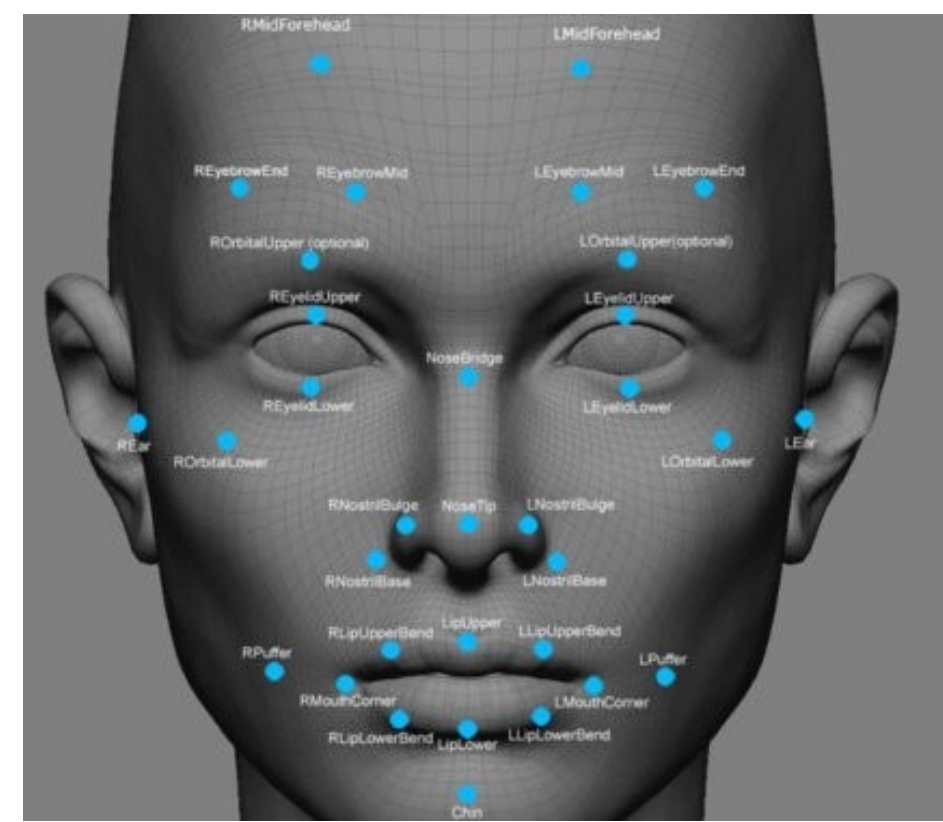


# Why classify AI systems depending on characteristics?

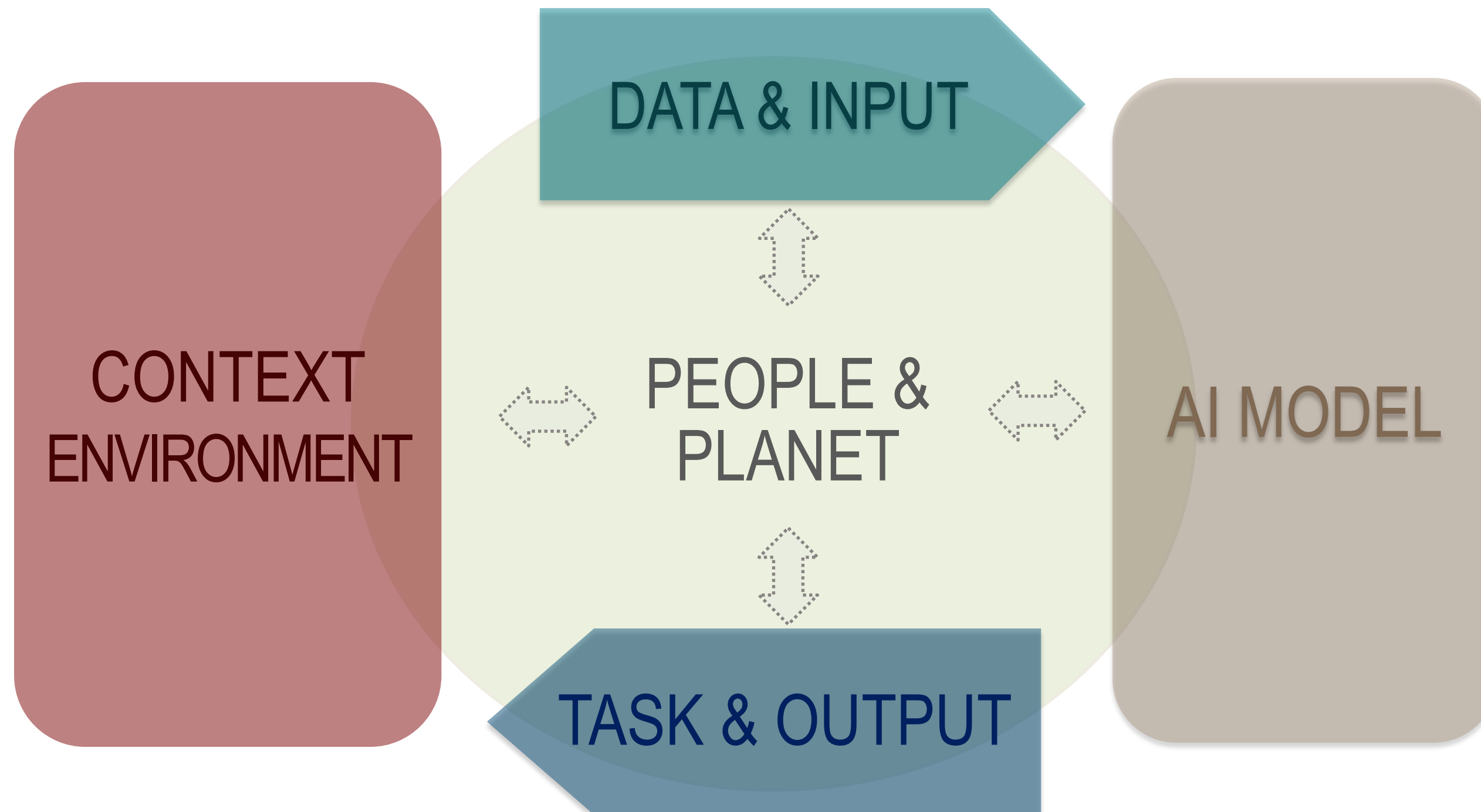
*A variety of systems and policy implications*



● Experimental result  
● Computational prediction



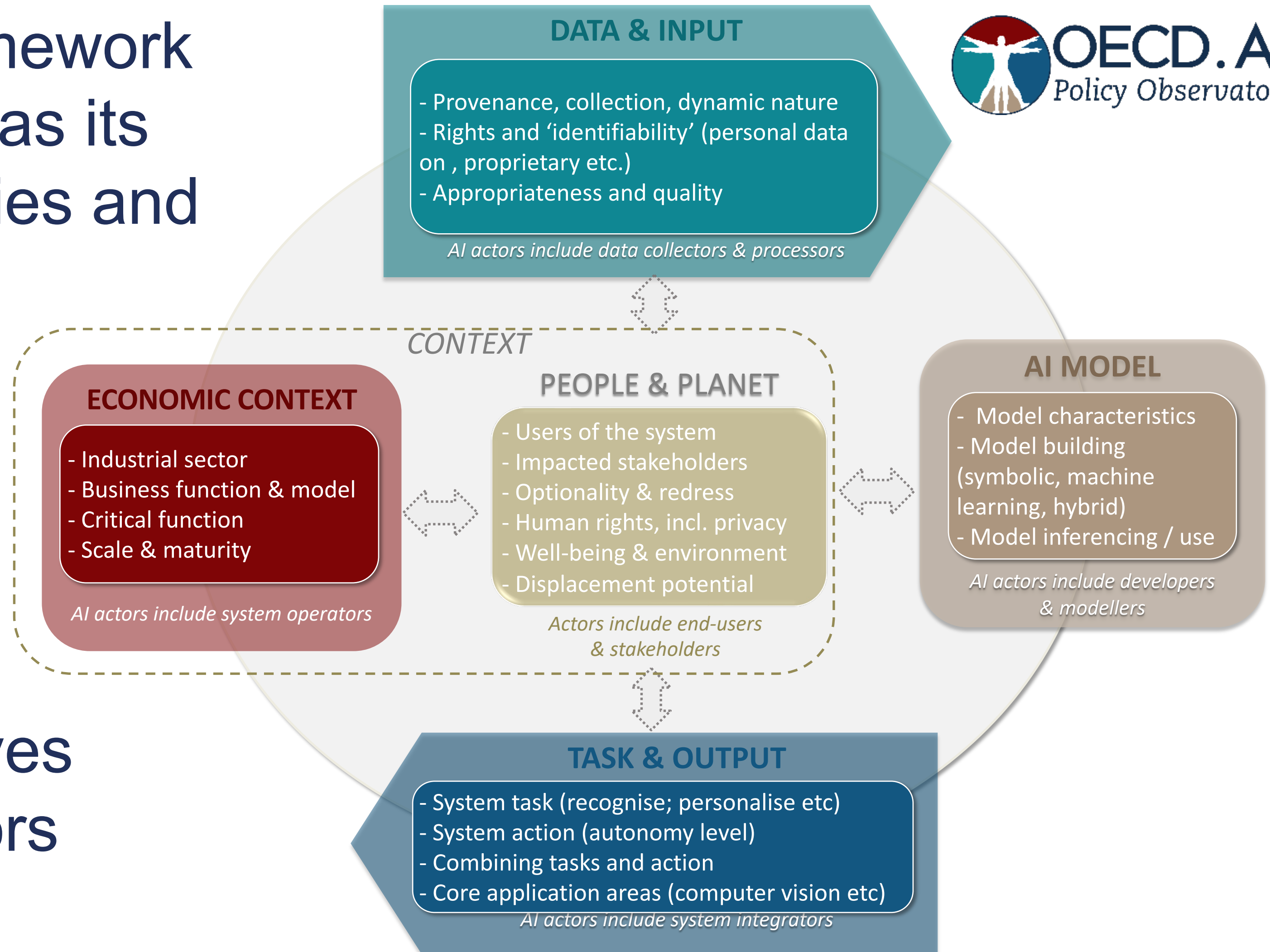
# The OECD Framework for Classifying AI systems: Key dimensions characterise AI systems' policy impact



For more see: [OECD.AI/classification](https://oecd.ai/classification)



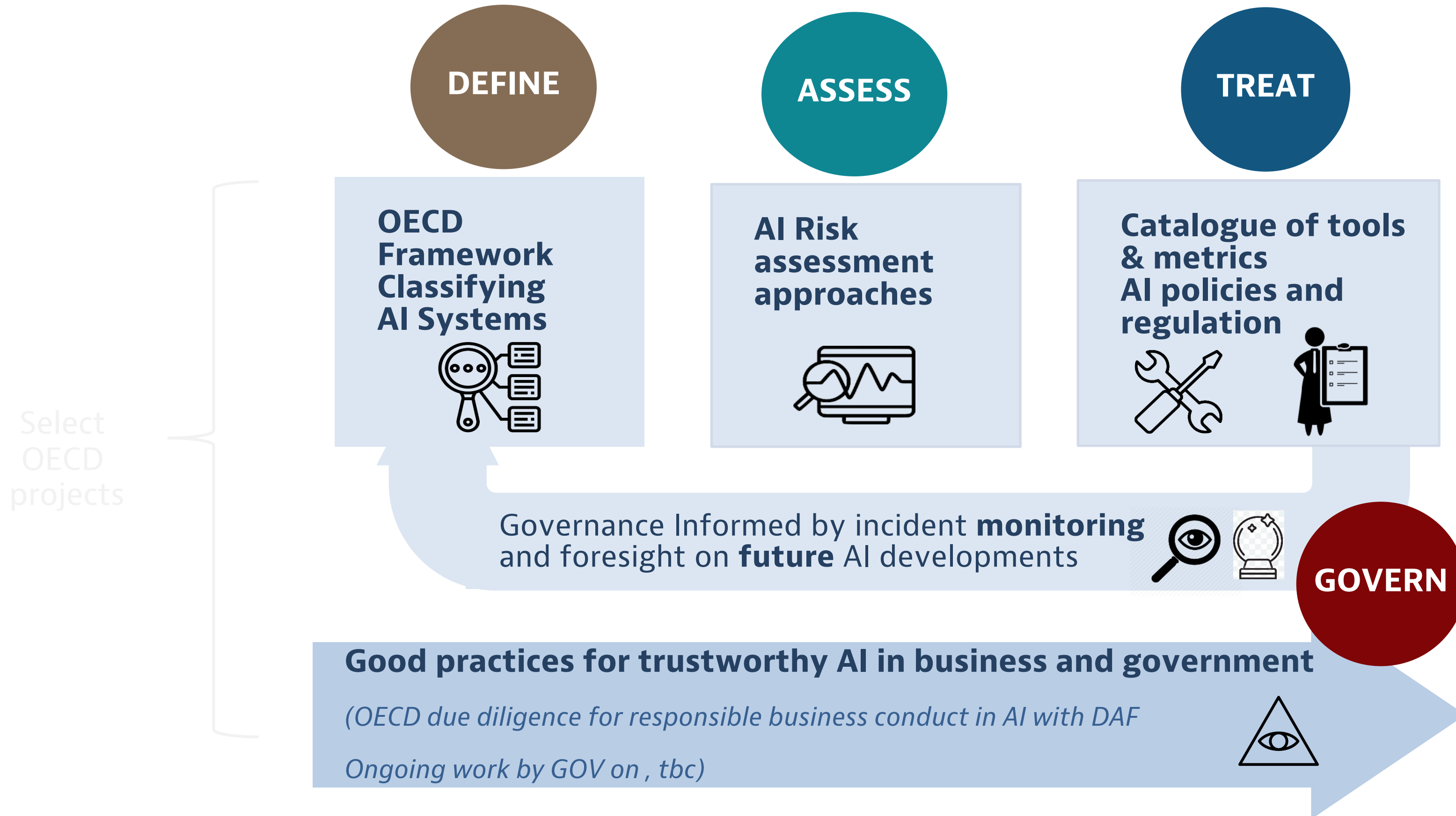
Each AI framework dimension has its own properties and attributes...



...and involves specific actors

# Implementing trustworthy, values-based AI

## OECD contributions to managing AI risks effectively



**Thank you**





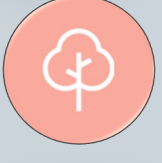
*For more information please visit*

**[www.oecd.ai](http://www.oecd.ai)**

**email: [ai@oecd.org](mailto:ai@oecd.org)**



# Helping to implement policy recommendations

5 recommendations for national policies	Select projects
 R&D	<ul style="list-style-type: none"><li>▪ Long-term AI developments</li><li>▪ Smart energy systems</li></ul>
 Data, compute & technologies	<ul style="list-style-type: none"><li>▪ AI compute and climate - AI language technologies</li><li>▪ Data for AI</li></ul>
 Enabling policies & regulations	<ul style="list-style-type: none"><li>▪ AI regulatory sandboxes</li><li>▪ National country reviews</li></ul>
 Jobs & skills	<ul style="list-style-type: none"><li>▪ Programme on AI &amp; work, innovation, productivity and skills</li></ul>
 Cooperate internationally	<ul style="list-style-type: none"><li>▪ Globalpolicy.ai (with 8 IGOs incl. EC, CoE, UNESCO, IDB)</li><li>▪ Global Partnership on AI (GPAI)</li></ul>

## SUPPORTING IMPLEMENTATION OF ALL POLICY RECOMMENDATIONS:

- Measuring trends re. policy recommendations
- Tracking policies & regulatory developments in 60+ countries (with EC)



# Upcoming Events



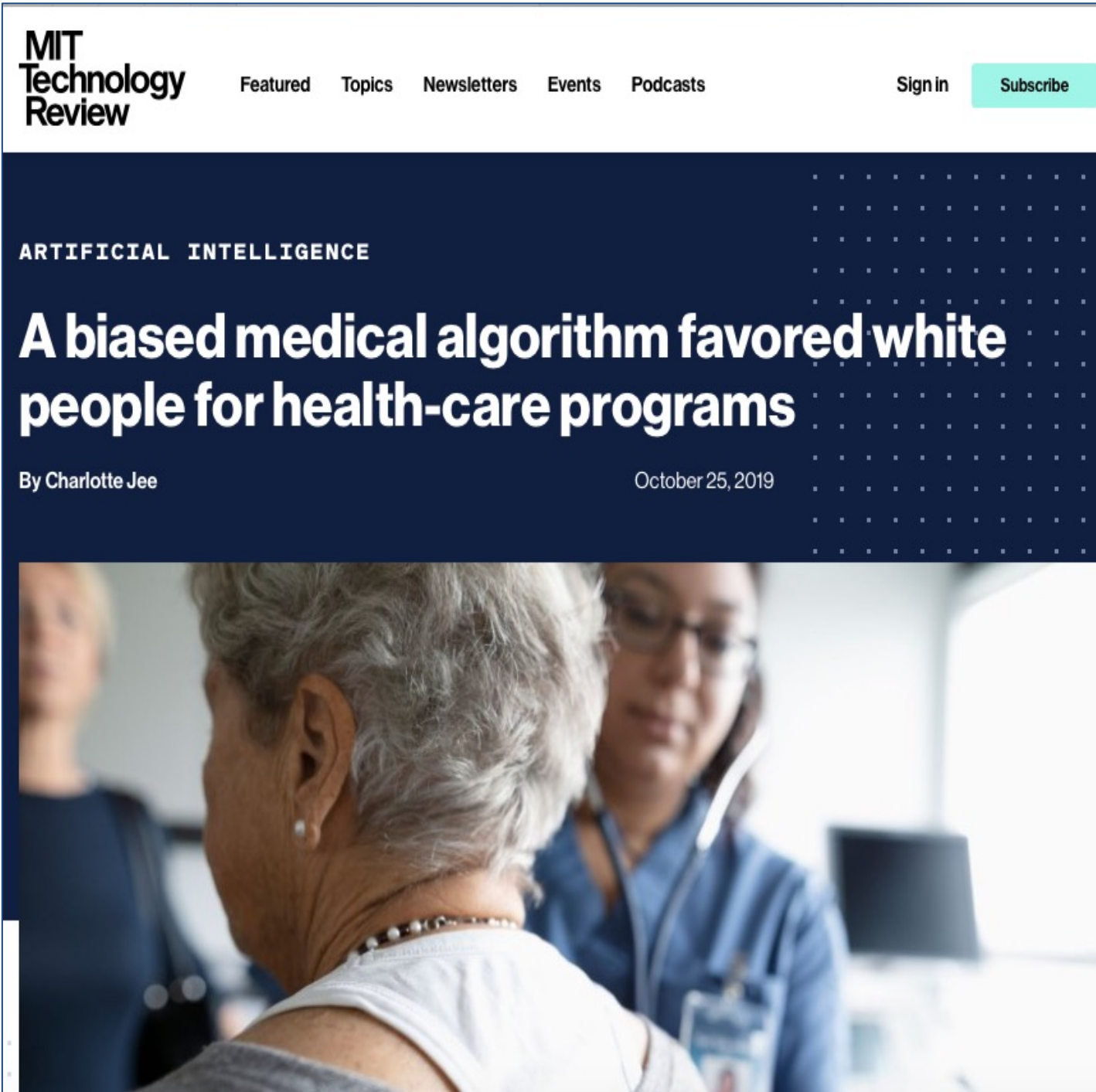
- **OECD Digital Economy Ministerial – 13-16 Dec 2022**
- Launch of the **catalogue of tools** for trustworthy AI – December 2022 or January 2023 (TBC) 
- Third Global Conference on AI and **work, innovation, productivity and skills** (AI-WIPS), supported by Germany, February 2023





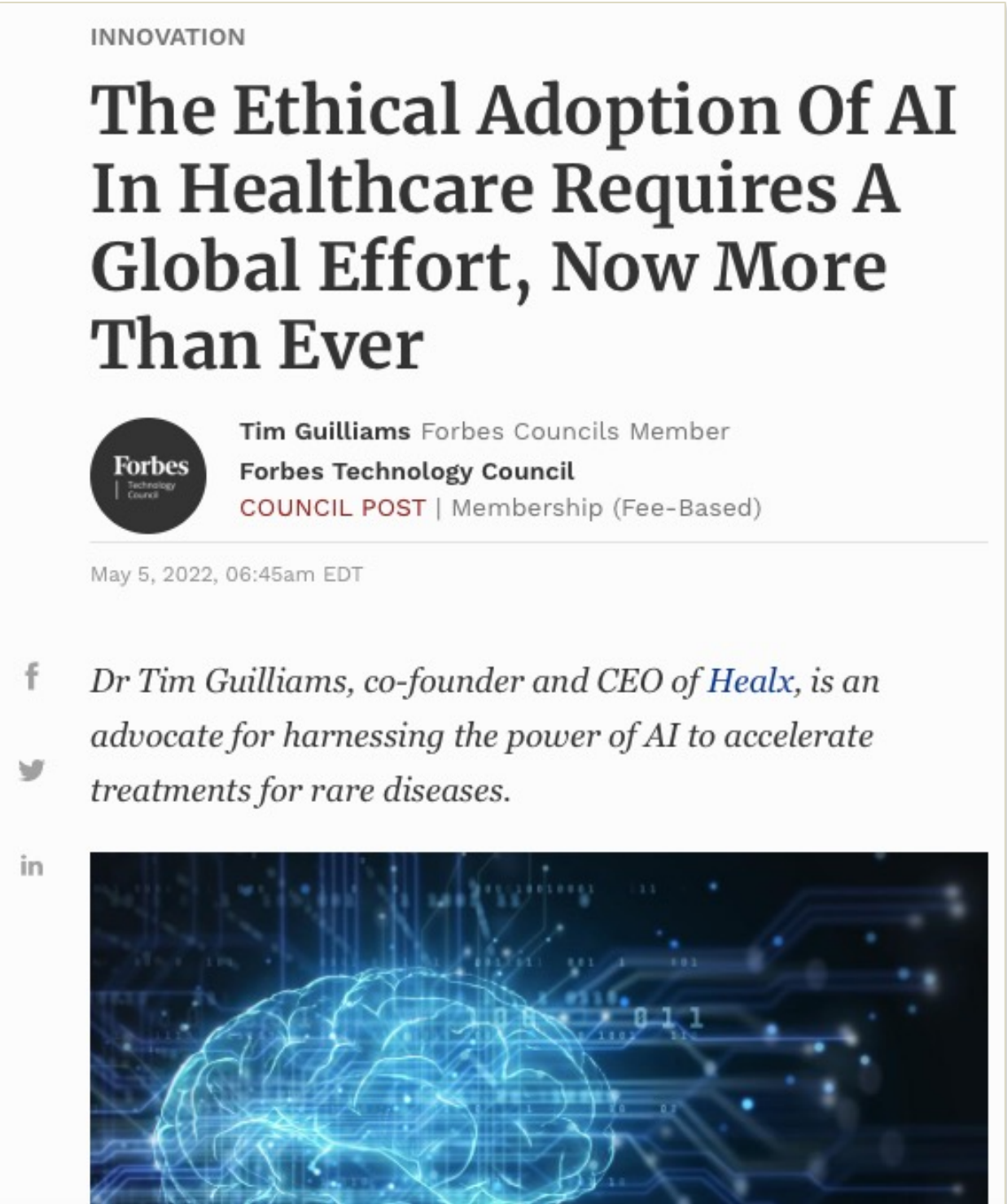
# Step 2: Automatic identification of incidents

Examples of automating AI incidents identification in real time using news articles




Probability: 0.3705  
NOT AI INCIDENT

Probability: 0.7305  
AI INCIDENT





# Catalogue of tools & metrics for trustworthy AI, with US NIST



OECD.AI  
Policy Observatory

OECD.orgGoing Digital ToolkitEN

Experts & blogAI PrinciplesPolicy areasTrends & dataTools catalogueCountriesAbout

HomeTools catalogue

Catalogue of tools for trustworthy AI

An interactive collection of the latest tools and resources to help AI actors be accountable and ensure that AI systems and applications respect human rights and are fair, transparent, explainable, robust, secure and safe.

TYPE

Approach

☐ Technical  
☐ Procedural  
☐ Educational  
☐ Other

Tool type

Filter by...

OBJECTIVE

Objective

Filter by...

ORIGIN

Stakeholder group

Country

Organisation

Filter by...

SCOPE

Lifecycle stage(s)

Target group(s)

Target user(s)

Target sector(s)

Impacted stakeholders

Application task(s)

ADOPTABILITY

Tool maturity

Licensing regime

List of tools63 tools found under the current selection

[LinkedIn Fairness Toolkit \(LiFT\)](#)

TechnicalUnited States

Open source toolkit to enable measurement of fairness according to a multitude of fairness definitions in large-scale machine learning workflows. LinkedIn Fairness Toolkit (LiFT), is an open source Scala/Spark library that enables the measurement of fairness, according to a multitude of fairness definitions, in large-scale machine learning workflows.

Objective(s)

Fair & unbiased

Related lifecycle stage(s)

Build & interpret model, Collect & process data, Deploy, Operate & monitor, Plan & design, Verify & validate

[Mozilla Open Source Audit Tooling \(OAT\) Project](#)

TechnicalUnited States

Over the coming year, Mozilla Fellow Deb Raji is running the Open Source Audit Tooling (OAT) Initiative. Deb will identify the resources and tools needed to support algorithmic auditors, and to make thorough and consequential AI scrutiny the status quo.

Objective(s)

AccountableRobust & secure  
Transparent & explainable

Related lifecycle stage(s)

Operate & monitor

[Microsoft InterpretML](#)

TechnicalUnited States

An open-source toolkit containing machine learning interpretability algorithms to help understand model predictions.

Objective(s)

AccountableTransparent & explainable

Related lifecycle stage(s)

Build & interpret model, Collect & process data, Plan & design, Verify & validate


[TOOLBOX: Dynamics of AI Principles](#)

Educational, ProceduralUnited States

AI Ethics Lab created the Dynamics of AI Principles to help understand the global trends, commonalities, and differences among numerous sets of AI principles published.

Related lifecycle stage(s)

All stages



OECD.AI  
Policy Observatory

OECD.orgGoing Digital ToolkitEN

Experts & blogAI PrinciplesPolicy areasTrends & dataTools catalogueCountriesAbout

HomeTools catalogueLinkedIn Fairness Toolkit (LiFT)

LinkedIn Fairness Toolkit (LiFT)

Open source toolkit to enable measurement of fairness according to a multitude of fairness definitions in large-scale machine learning workflows. LinkedIn Fairness Toolkit (LiFT), is an open source Scala/Spark library that enables the measurement of fairness, according to a multitude of fairness definitions, in large-scale machine learning workflows.

WebsiteGithubSlack

TechnicalTools/softwareUnited Statespublished on Mar 1, 2022

Organisation(s): LinkedIn

Open source toolkit to enable measurement of fairness according to a multitude of fairness definitions in large-scale machine learning workflows. The LinkedIn Fairness Toolkit (LiFT), is an open source Scala/Spark library that enables the measurement of fairness, according to a multitude of fairness definitions, in large-scale machine learning workflows.

The LinkedIn Fairness Toolkit library has broad utility for organizations who wish to conduct regular analyses of the fairness of their own models and data.


It can be deployed in training and scoring workflows to measure biases in training data, evaluate different fairness notions for ML models, and detect statistically significant differences in their performance across different subgroups. It can also be used for ad hoc fairness analysis or as part of a large-scale A/B testing system.

Current metrics supported measure different kinds of distances between observed and expected probability distributions, traditional fairness metrics (e.g., demographic parity, equalized odds), and fairness measures that capture a notion of skew like Generalized Entropy Index, Theil's Index, and Atkinson's Index.

LiFT also introduces a novel metric-agnostic permutation testing framework that detects statistically significant differences in model performance (as measured according to any given assessment metric) across different subgroups. This [testing methodology](#) will appear at [NIPS 2020](#).

[Read less](#)

Post about this tool



Will LinkedIn's Fairness Toolkit Mark the End of AI Bias?  
March 1, 2022 — 1 min read

About the tool

You can click on the link type items to see the associated tool

Countries:United States

Lifecycle(s) stage:Build & interpret model  
Collect & process data  
Deploy  
Operate & monitor  
Plan & design  
Verify & validate

Type of approach:Technical

Objectives:Fair & unbiased

Organisation:LinkedIn

Maturity:Implemented in multiple projects

Licensing regime:Open source


Target sectors:All

Target users:Business leader  
Data scientist  
Developer  
System operators


Stakeholder group:Business

Impacted stakeholders:Consumers  
Employees


Use Cases



Finance UK  
**LiFT for SMEs in the financial industry**  
What is Lorem Ipsum? Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and



Analytics Drift  
**LinkedIn Fairness Toolkit (LiFT) For Explainability in Machine Learning**  
By Analytics Drift LinkedIn Fairness Toolkit (LiFT) was released by the largest professional networking giant to



LinkedIn  
**Using the LinkedIn Fairness Toolkit in large-scale AI systems**  
By Praveen Nandy Introduction LinkedIn's vision to create economic opportunity for every member of the global

# Towards interoperable AI incident reporting frameworks

- What constitutes an “AI incident”?
- What scope of incidents should be covered?
- Lessons from other areas?

WHAT?

WHO?

- Who should report?
- To whom should reports be made?
- In which circumstances?

Key  
questions

WHY?

HOW?

- What *incentives/ rewards* to share information? What *barriers*?
- Private, full, responsible or coordinated disclosure?
- Voluntary or mandatory? What does voluntary mean?
- Under what conditions?
- What role for trust?

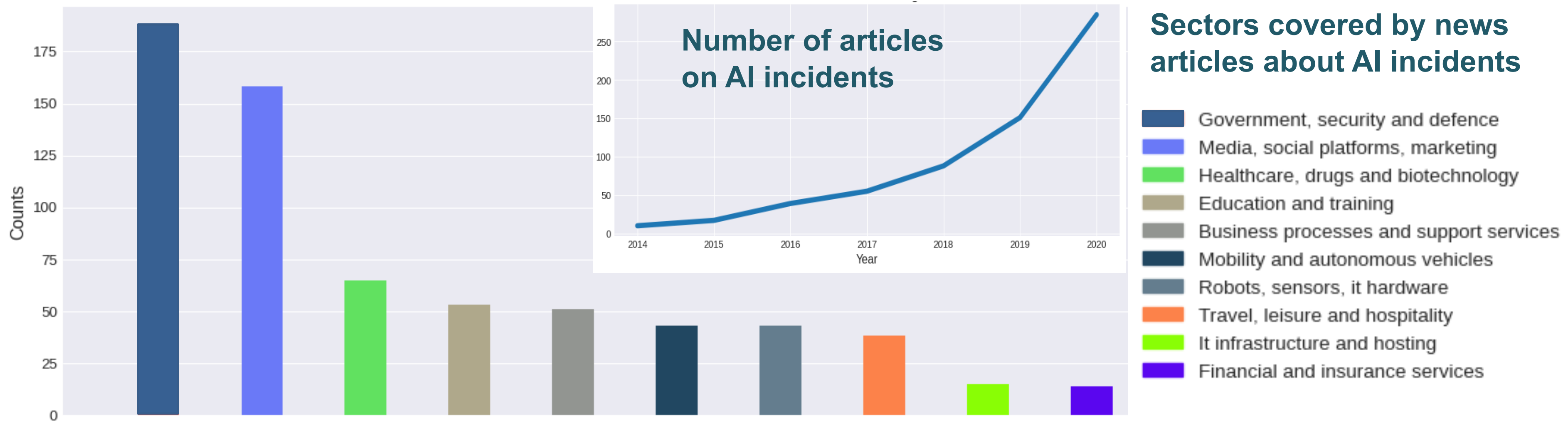
- What information should be shared with who?
- With selected actors? publically? governments?...



# Step 1: Collection of training data

Illustrative findings from manually identified news articles on AI incidents and hazards

*Caveat:* significant sampling bias (not the whole story)



Source: McGregor, S. (2021) Preventing Repeated Real World AI Failures by Cataloging Incidents: [The AI Incident Database](#). In Proceedings of the Thirty-Third Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-21). Virtual Conference.; [AIAAIC's incident and controversy repository](#); AI Global's [map of responsible and harmful AI](#).



# Responsible Business Conduct (RBC) for Trustworthy AI

- Businesses need consistent international guidance endorsed by governments to help manage AI risks
- RBC guidelines well-established internationally (50 countries)
  - Cover numerous areas (e.g. minerals, agriculture and finance)
  - Companies of all sizes that operate internationally
  - Include human rights (HR) due diligence, complementary to legislation
- Enforcement mechanisms / dispute resolution already in place

**RBC for AI: high returns for minimal effort.  
Effective option for accountability  
without stifling innovation in a fast-moving area like AI.**



DEFINE

ASSESS

TREAT

GOVERN

High level AI risk management interoperability framework

