

# *1. OECD work to operationalise trustworthy AI*

**Andrew Wyckoff**

**High-Level Roundtable on Artificial Intelligence:  
Regulating Innovation and Innovating Regulation**

*22 March 2022*

# OECD AI Principles

**10 Principles, covering two areas:**

## **Principles for responsible stewardship of trustworthy AI**

-  Inclusive growth, sustainable development and well-being
-  Human-centred values and fairness
-  Transparency and explainability
-  Robustness, security and safety
-  Accountability

## **National policies and international cooperation for trustworthy AI**

-  Investing in AI research and development
-  Fostering a digital ecosystem for AI
-  Providing an enabling policy environment for AI
-  Building human capacity and preparing for labour transition
-  International cooperation

## OECD AI Policy Observatory (OECD.AI)

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

### 5 pillars:

- Network of experts and AI Wonk blog
- AI Principles & implementation
- AI trends & data
- AI policy areas
- Countries & initiatives



## OECD Working Party on AI Governance and AI Experts Groups

*Developing practical guidance to implement the AI Principles.*

### 1 formal working party, 3 expert groups:

- OECD Working Party on AI Governance (AIGO)

Supported by:

- Expert Group on AI Classification & Risk
- Expert Group on AI Tools & Accountability
- Task force on AI compute

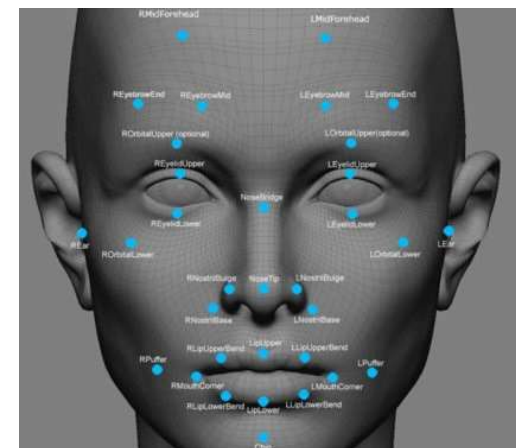
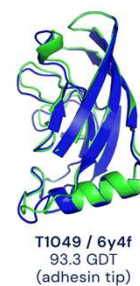
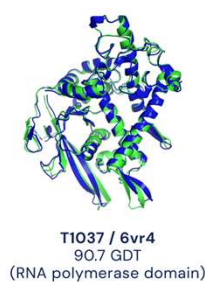
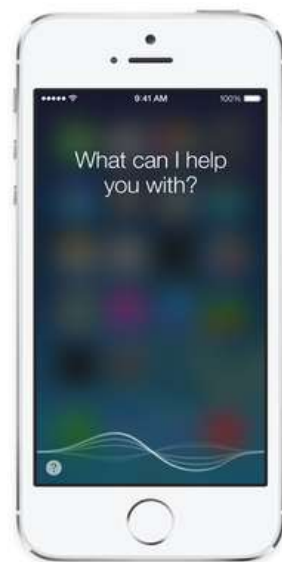
### In addition to:

- The OECD Global Parliamentary Group on AI
- The Global Partnership on AI (GPAI)

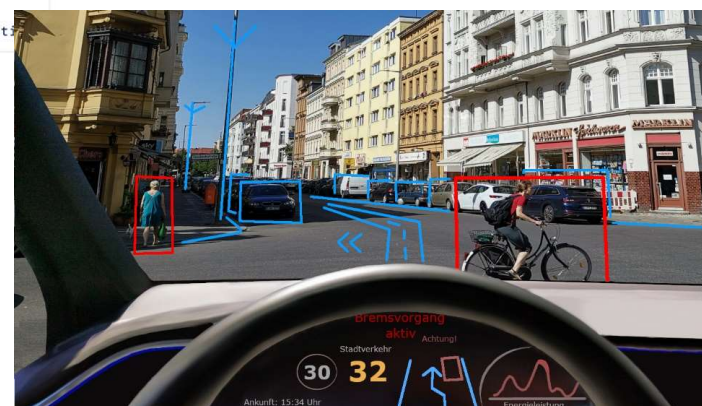
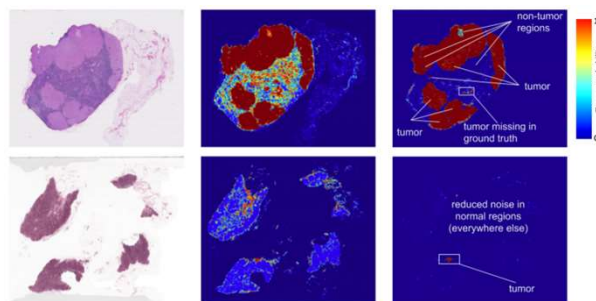


# The OECD Framework for Classifying AI Systems

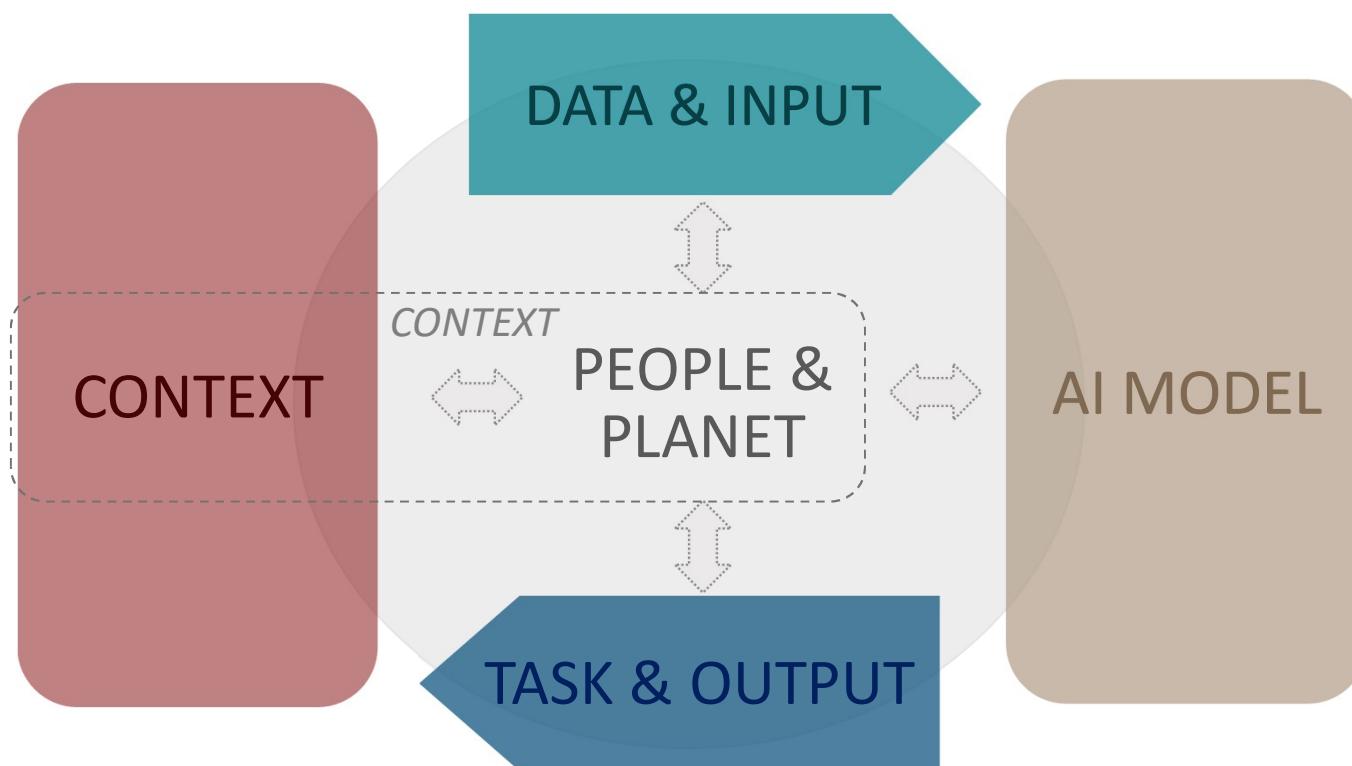
## Why classify AI systems? *A variety of systems and policy implications*



● Experimental result  
● Computational prediction

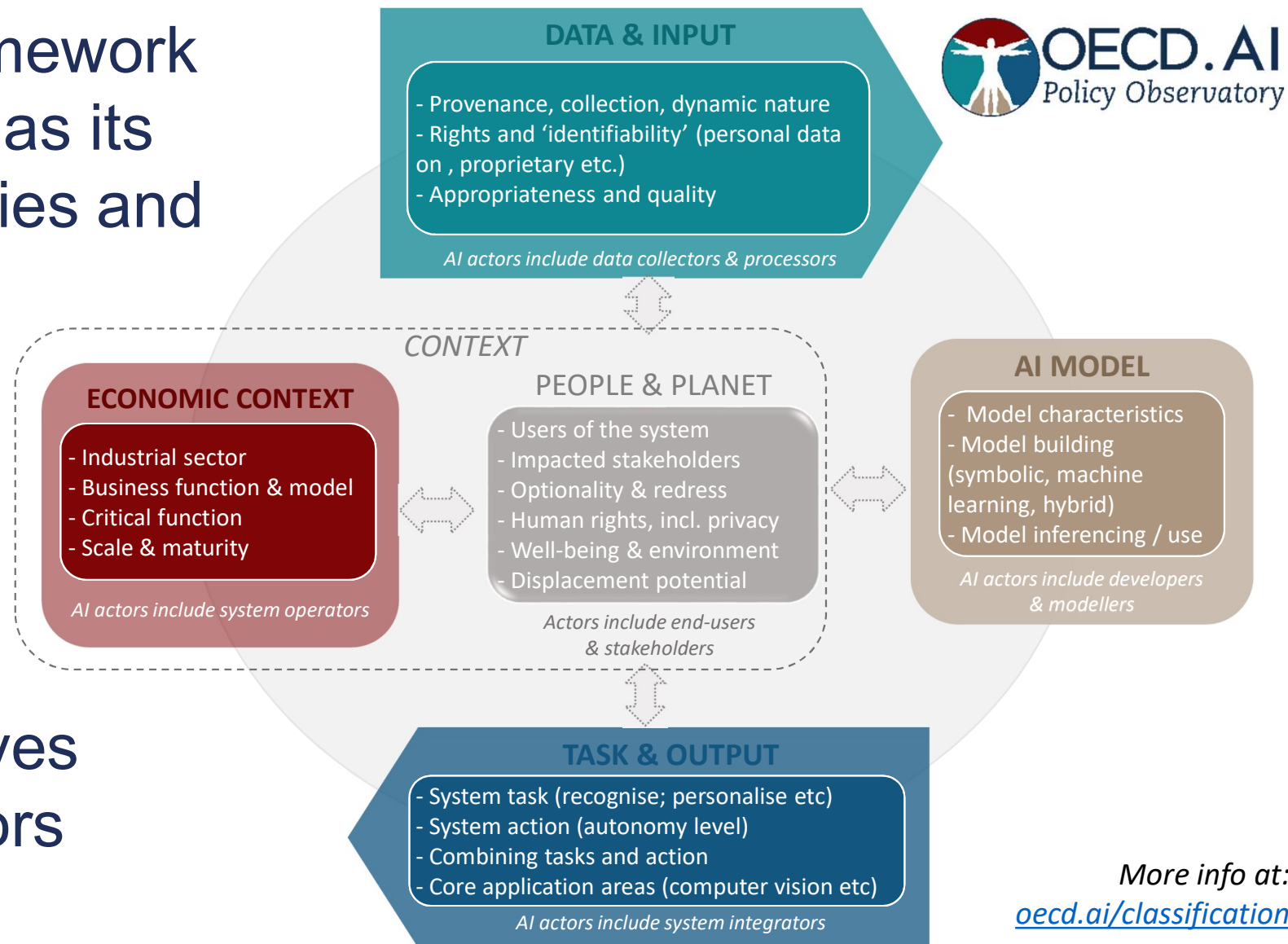


# OECD Framework for Classifying AI systems: Key dimensions characterise AI systems' policy impact



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

...and involves specific actors



## Next steps at the OECD:



- **Refine classification criteria**
  - Add more real-world AI systems and identify possible indicators
- **Develop a risk assessment framework to facilitate global interoperability**
  - Leveraging work in partner organisations, including EU, US, ISO
  - Leveraging risk assessment work in other parts of the OECD
- **Develop a common framework for reporting about AI incidents**
- **Support risk management:** Inform related work on mitigation, compliance and enforcement along the AI system lifecycle, and responsible business-impact assessment.



# Catalogue of tools for trustworthy AI

## Catalogue of tools for trustworthy AI

**OECD AI Policy Observatory**

Experts & blog | AI Principles | Policy areas | Trends & data | **Tools catalogue** | Countries | About | Q

Home > Tools 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 tools** 63 tools found under the current selection

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

Technical United 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](#)

Technical United 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)** Accountable, Robust & secure, Transparent & explainable

**Related lifecycle stage(s)** Operate & monitor

#### [Microsoft InterpretML](#)

Technical United States

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

**Objective(s)** Accountable, Transparent & explainable

**Related lifecycle stage(s)** Build & interpret model, Collect & process data, Plan & design, Verify & validate

#### [TOOLBOX: Dynamics of AI Principles](#)

Educational, Procedural United 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**

Experts & blog | AI Principles | Policy areas | Trends & data | **Tools catalogue** | Countries | About | Q

Home > Tools catalogue > **LinkedIn 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.

Website | GitHub | News

Technical United States published on Mar 1, 2024

**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 organisations 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 metrics 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 AIS testing system.
- Current metrics supported measure different kinds of distances between observed and expected probability distributions, traditional fairness metrics (e.g., demographic parity, equalised odds), and fairness measures that capture a notion of view like Fairness-aware Causality Index, Pearl's Inference, and Algorithm'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 in [IEEE 2024](#).

[Read less](#)

**Post about this tool**

[Will LinkedIn's Fairness Toolkit Mark the End of AI Bias?](#)

March 1, 2024 | 15 min read

**About the tool**

You can click on the tool type items to see the associated tools.

**Countries:** [United States](#)

**Lifecycle stage(s):** [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 leaders](#), [Data scientists](#), [Developers](#), [System architects](#)

**Stakeholder group:** [Business](#)

**Impacted stakeholders:** [Consumers](#), [Employees](#)


### Use Cases



**Finance UK**

**LiFT for SMEs in the financial industry**


What is Laven (Spain)? Laven (Spain) is simply dummy text of the printing and typesetting industry. Laven (Spain) has been the industry's standard dummy text ever since the 1500s.



**Analytics Delft**

**LinkedIn Fairness Toolkit (LiFT) For Explainability in Machine Learning**

By Analytics Delft LinkedIn Fairness Toolkit (LiFT) was released by the largest professional networking giant to enhance explainability in machine learning.



**LinkedIn**

**Using the LinkedIn Fairness Toolkit in large-scale AI systems**

By Pearson Ready Introduction LinkedIn's vision to create economic opportunity for every member of the global



# Analytical work on auditing AI systems

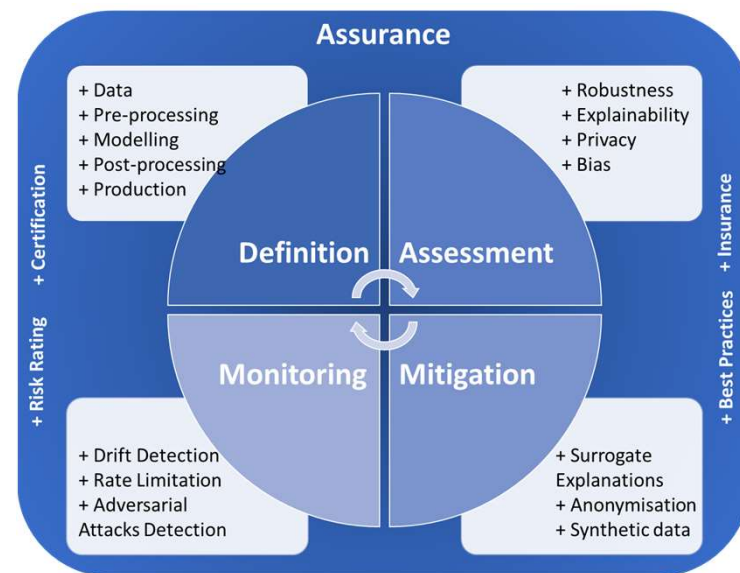


## Next step: encouraging international interoperability in risk assessment



At each stage of the lifecycle conduct, conduct a risk assessment “DAMMA”:

1. **DEFINE:** relevant principles for that stage, and relevant stakeholders and actors
2. **ASSESS:** risks to principles at individual and aggregate/societal levels (i.e., many small risks can amount to a big risk)
3. **MITIGATE:** in a way that is appropriate and commensurate to risk, considering likelihood and impact of risk
4. **MONITOR:** measure, evaluate and feedback results of the implementation
5. **ASSURE:** verify (audit, certify etc.) and communicate.



*For more information visit*  
[www.oecd.ai](http://www.oecd.ai)

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