OECD & TRUSTWORTHY AI: From Principles to Practice

6 December 2022
Lisbon Council
Karine Perset

OECD

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)

Other OECD committees and bodies

OECD Working Party on AI Governance (AIGO)

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

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

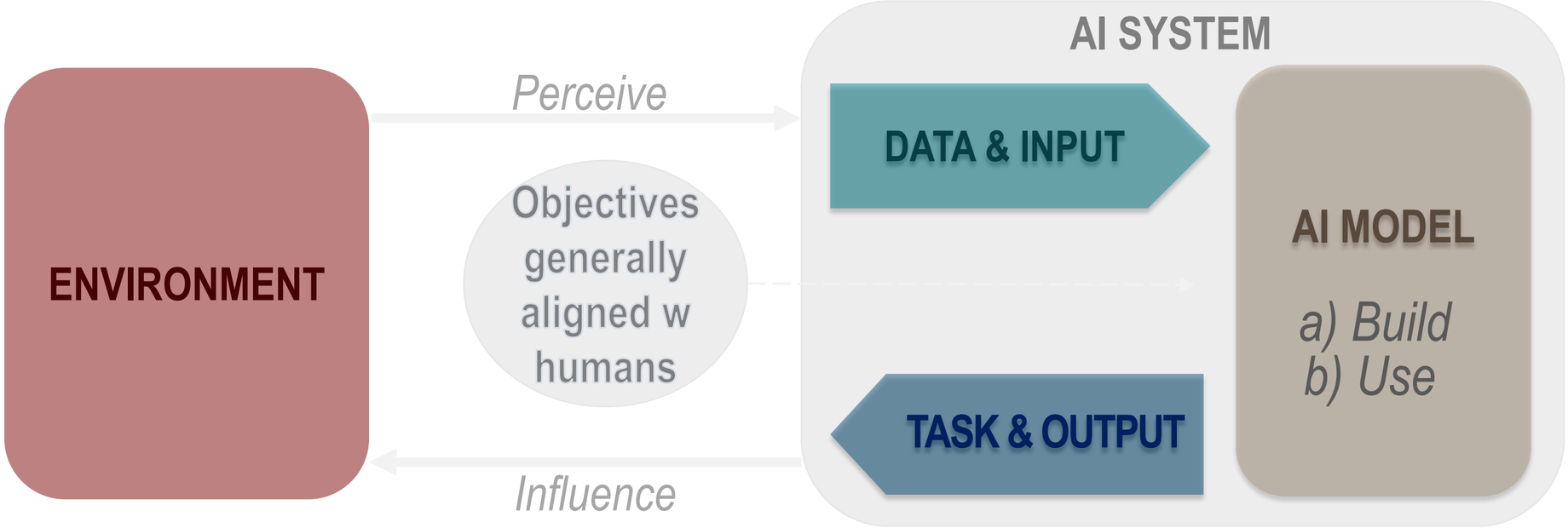
Expert group on AI foresight

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.

<u>WHAT IS AI</u>?... An AI system is a **machine-based** system that can...



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

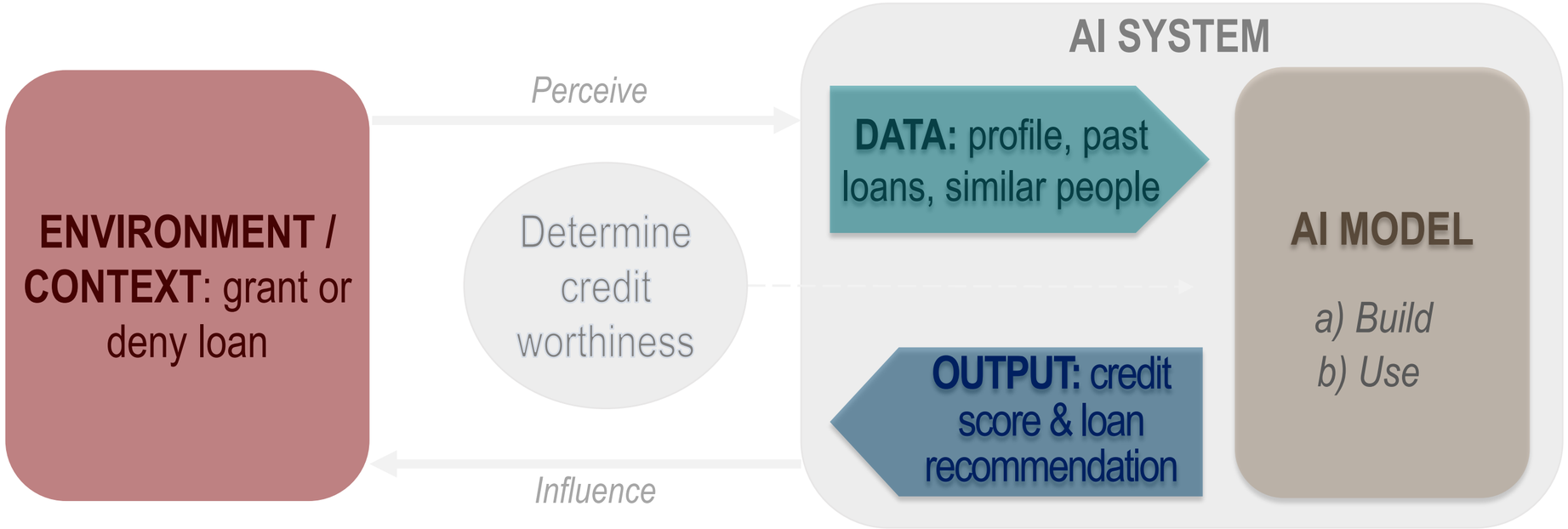
- perceive the environment using **data & inputs;**

- 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…

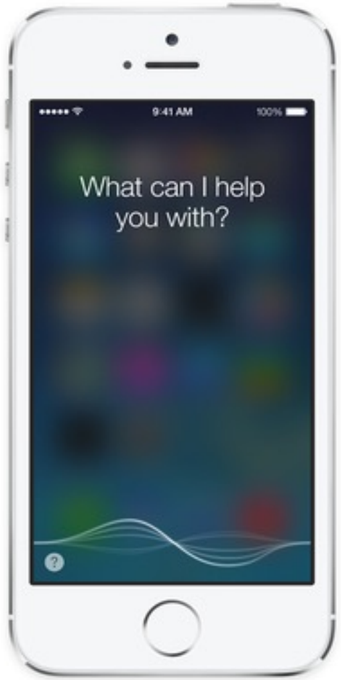


for a human-defined **objective** of determining credit-worthiness:

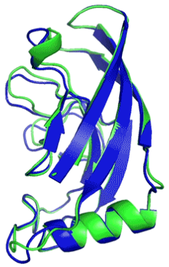
- perceive context using **data** on people, their past loans & similar people;
- abstract these perceptions to **build** a statistical credit-scoring model;
- **use the model** to generate a credit score and loan recommendation output
- that **influences** whether someone is granted or denied a loan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

Why classify AI systems depending on characteristics?

A variety of systems and policy implications

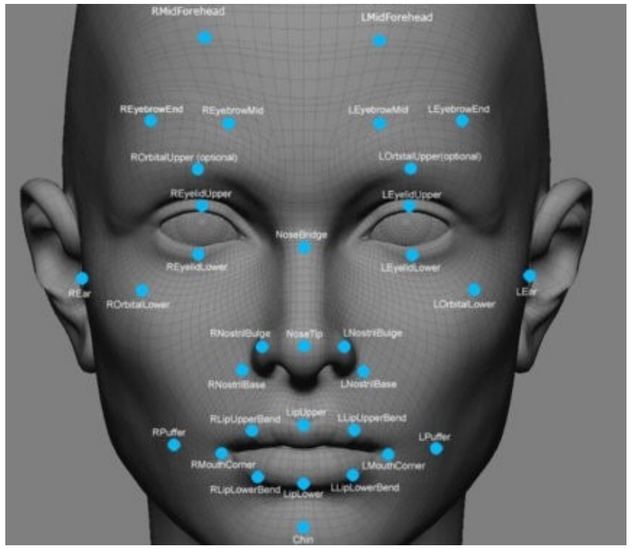


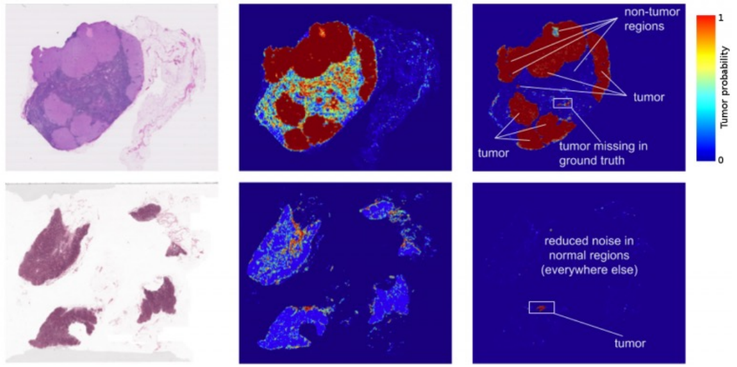
T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

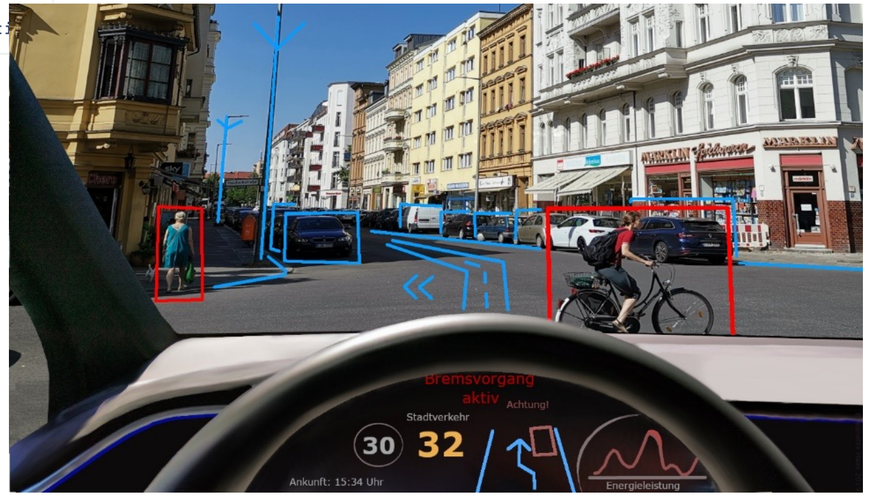


T1049 / 6y4f
93.3 GDT
(adhesin tip)

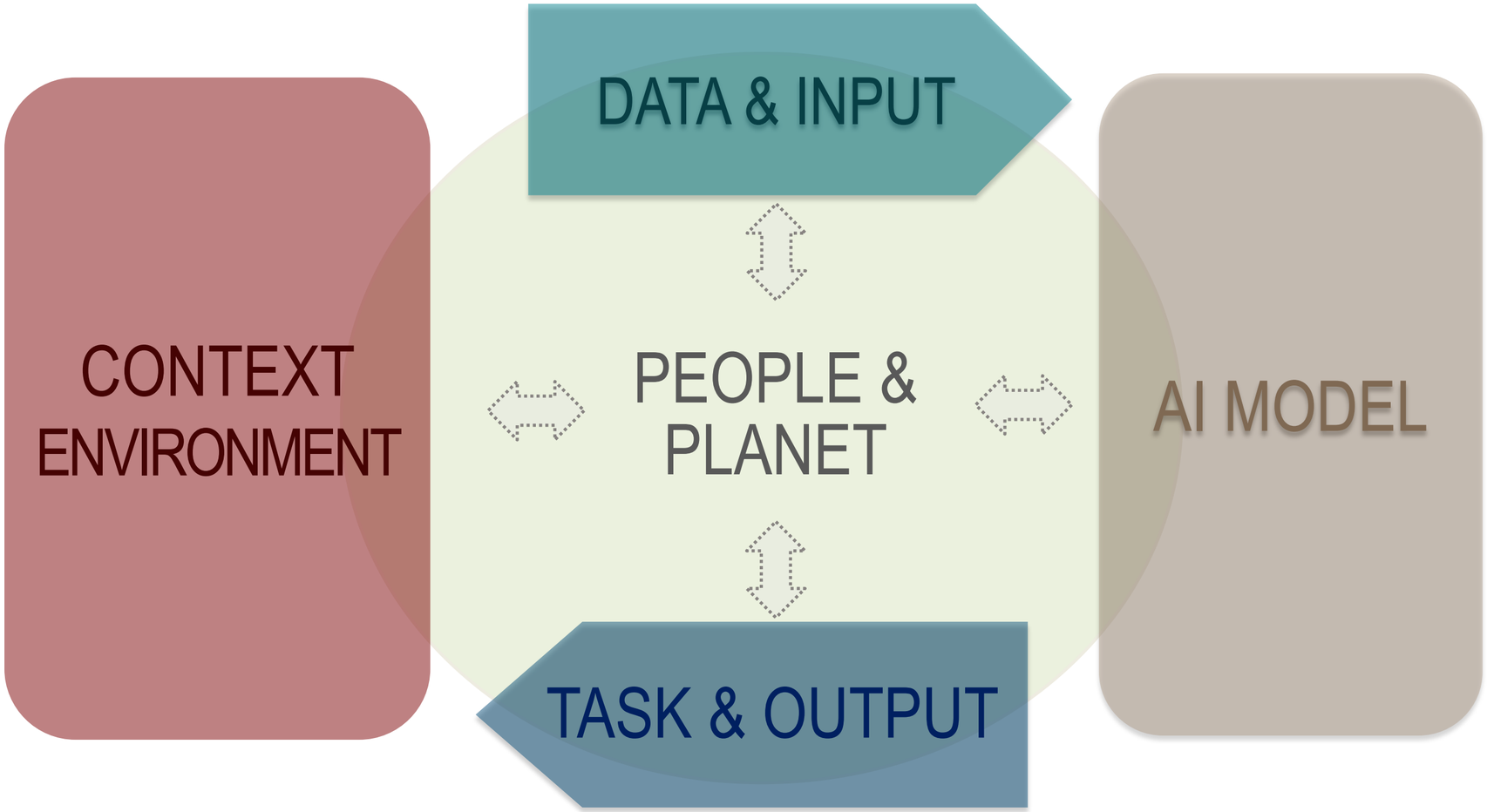
Experimental result
Computational predicti





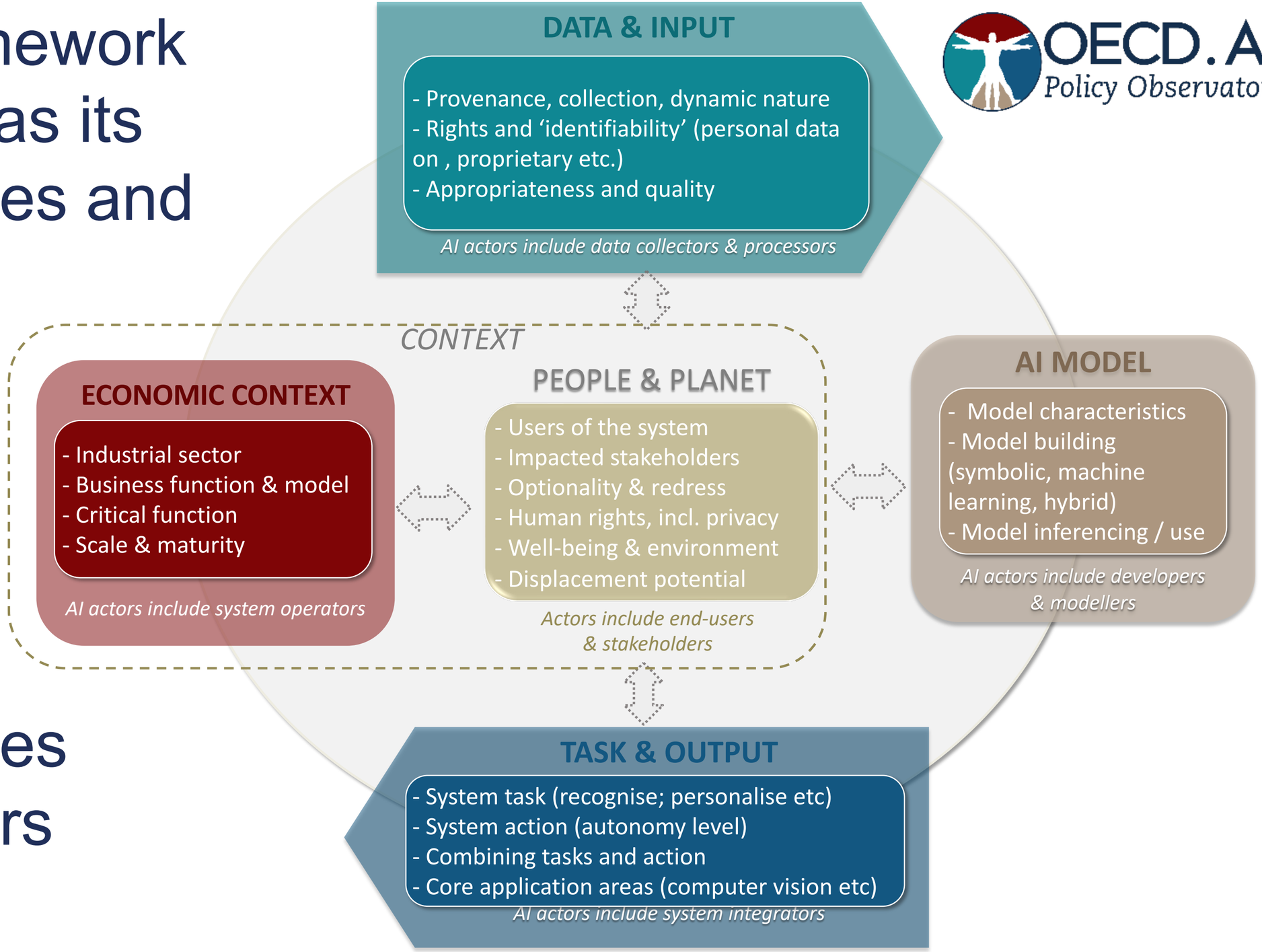


The OECD Framework for Classifying AI systems: Key dimensions characterise AI systems' policy impact



For more see: 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

Select OECD projects

DEFINE

OECD Framework Classifying AI Systems

ASSESS

AI Risk assessment approaches

TREAT

Catalogue of tools & metrics
AI policies and regulation

Governance Informed by incident **monitoring** and foresight on **future** AI developments

GOVERN

Good practices for trustworthy AI in business and government

(OECD due diligence for responsible business conduct in AI with DAF

Ongoing work by GOV on , tbc)

OECD.AI
Policy Observatory

Thank you

For more information please visit
www.oecd.ai

email: ai@oecd.org

Helping to implement policy recommendations

5 recommendations for national policies	Select projects
R&D	· Long-term AI developments · Smart energy systems
Data, compute & technologies	· AI compute and climate - AI language technologies · Data for AI
Enabling policies & regulations	· AI regulatory sandboxes · National country reviews
Jobs & skills	· Programme on AI & work, innovation, productivity and skills
Cooperate internationally	· Globalpolicy.ai (with 8 IGOs incl. EC, CoE, UNESCO, IDB) · Global Partnership on AI (GPAI)

SUPPORTING IMPLEMENTATION OF ALL POLICY RECOMMENDATIONS:

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

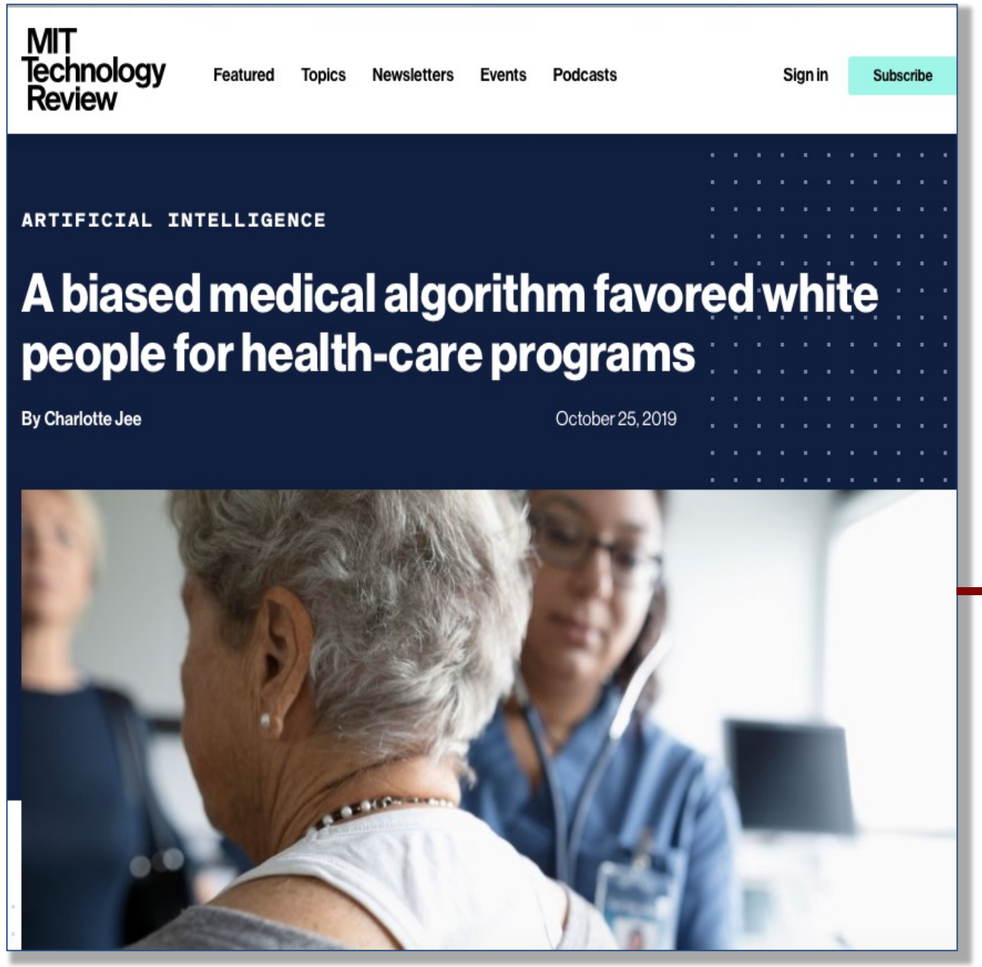
OECD.AI Policy Observatory

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

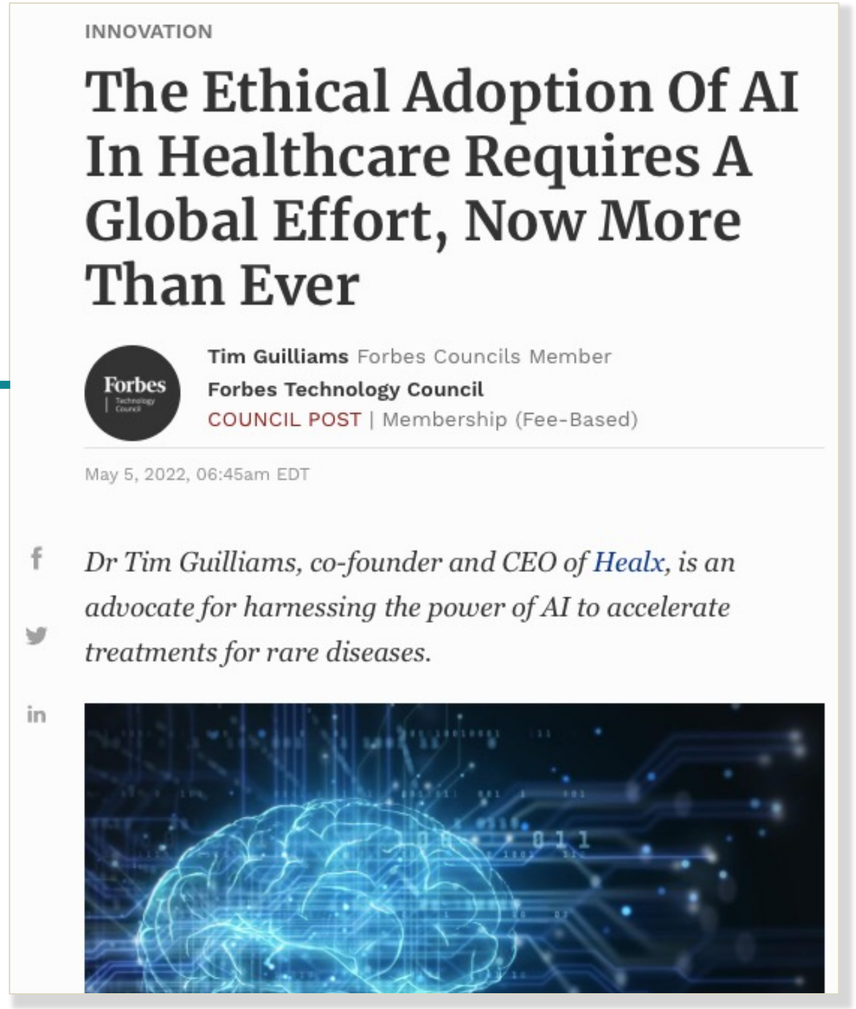
Step 2: Automatic identification of incidents

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

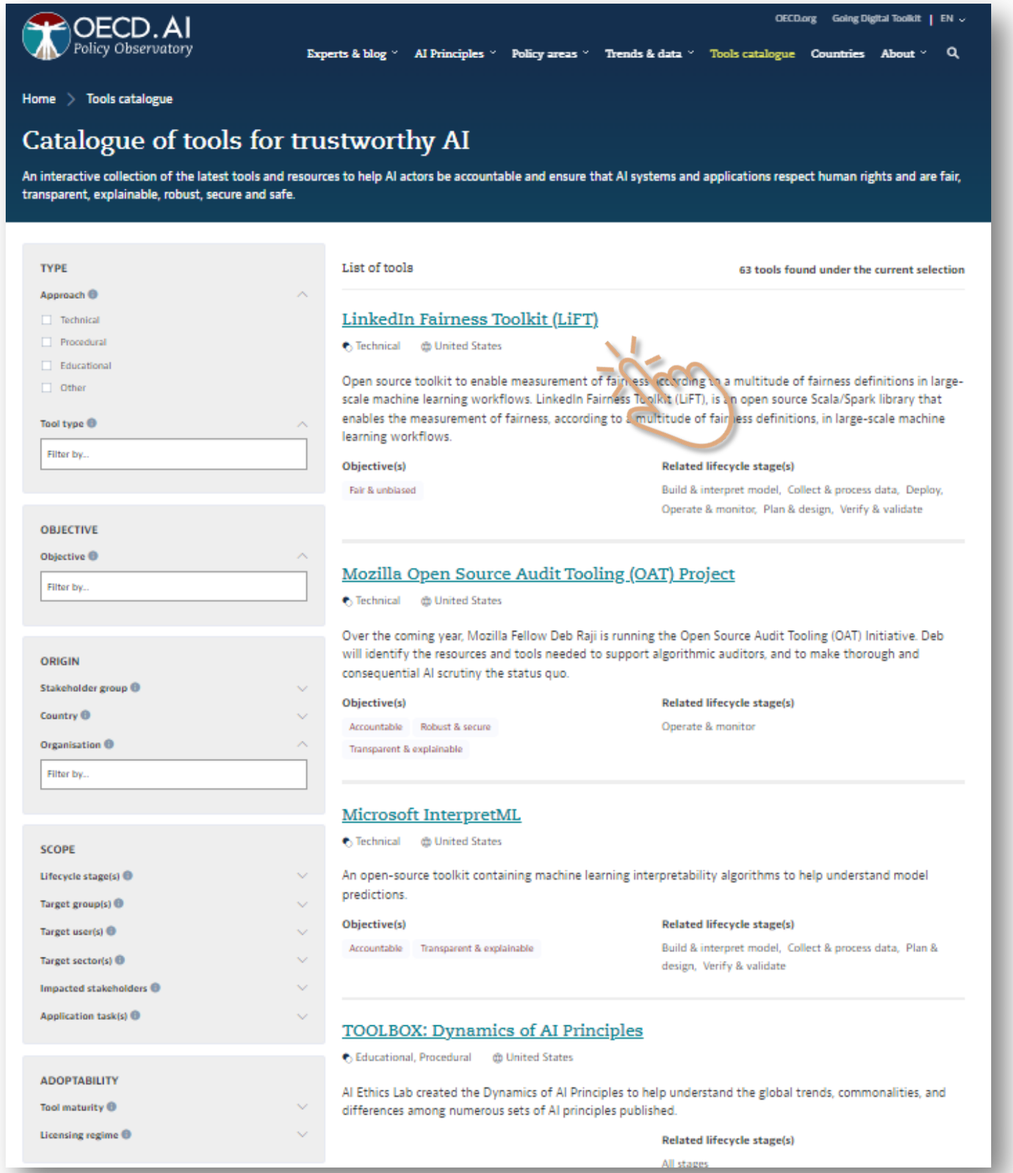


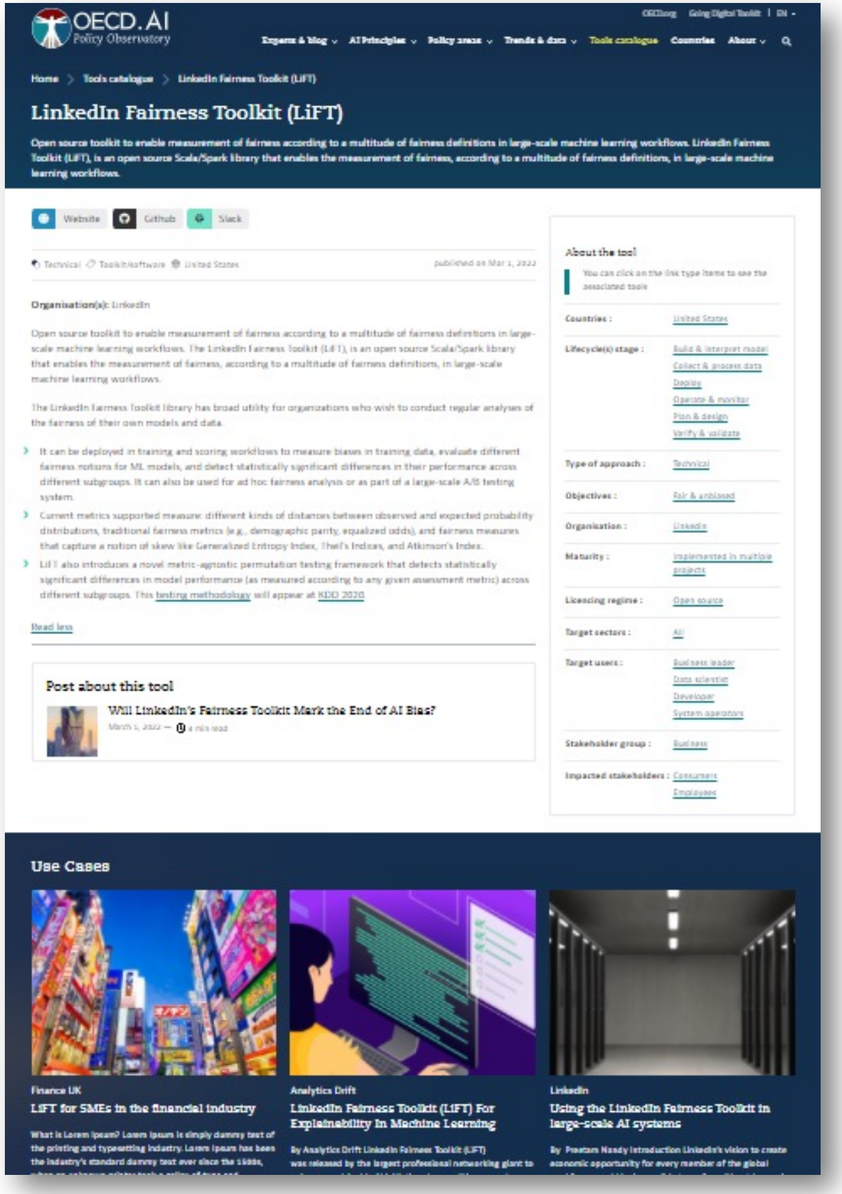
Probability: 0.7305
AI INCIDENT

Probability: 0.3705
NOT AI INCIDENT



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

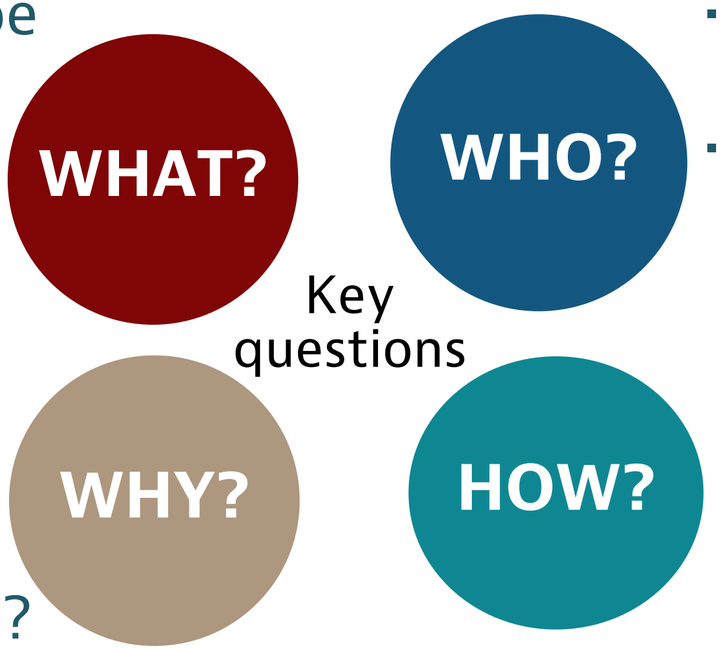




Towards interoperable AI incident reporting frameworks

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

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

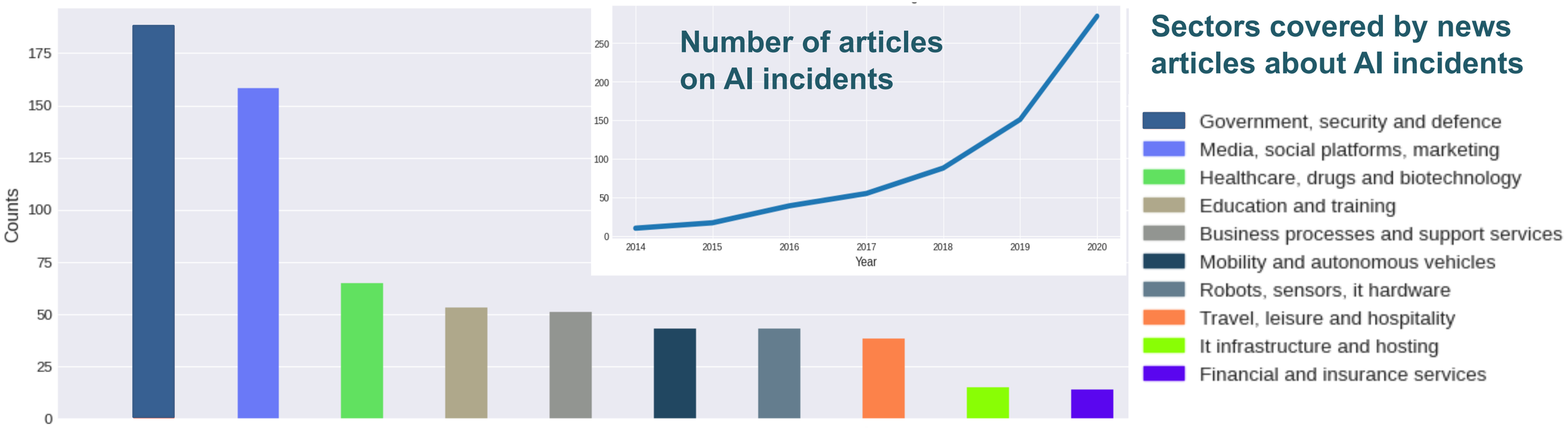


- What *incentives/ rewards* to share information? What *barriers*?
- Private, full, responsible or coordinated disclosure?
- Voluntary or mandatory? What does voluntary mean?
- Under what conditions?
- What role for trust?

- 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.

High level AI risk management interoperability framework