Enabling AI governance

OECD’s work on AI with a focus on robustness, security and safety

July 24, 2022
Vienna, Austria

Luis Aranda
Economist
OECD.AI
Outline

1. Background: The OECD AI Principles and OECD.AI
2. AI classification, risk and incidents
3. Catalogue of Tools for Trustworthy AI
The OECD AI Principles and OECD.AI
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
Principle 1.4: Robustness, security and safety

- AI systems should be robust, secure and safe throughout their entire lifecycle so that, in conditions of normal use, foreseeable use or misuse, or other adverse conditions, they function appropriately and do not pose unreasonable safety risk.

- To this end, AI actors should ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle, to enable analysis of the AI system’s outcomes and responses to inquiry, appropriate to the context and consistent with the state of art.

- AI actors should, based on their roles, the context, and their ability to act, apply a systematic risk management approach to each phase of the AI system lifecycle on a continuous basis to address risks related to AI systems, including privacy, digital security, safety and bias.
Governments that have committed to the OECD.AI Principles

OECD members

Adherents

G20 principles, based on OECD

*Singapore is an adherent
Moving from Principles to practice
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 policy areas
- AI trends & data
- Countries & initiatives

OECD Working Party on AI and Network of Experts

From Principles to practice…
National AI policies & strategies

This section provides a live repository of over 700 AI policy initiatives from 60 countries, territories and the EU. Click on a country/territory, a policy instrument or a group targeted by the policy.
AI classification, risk and incidents
5 interrelated work strands:

I. Develop a **framework for classifying AI systems** [completed]
   - To help policy-makers, regulators, legislators and others to assess the opportunities and risks that different types of AI systems present.

II. Develop a **risk assessment tool** [starting]
   - Leveraging ongoing initiatives including in the EU, US, ISO, IEEE.
   - Leveraging risk assessment work in other parts of the OECD.

III. Develop an **AI incident reporting framework** for global interoperability [in planning]
   - Framework for mandatory and voluntary reporting about AI incidents to ensure interoperability.
   - Build on lessons learnt from OECD Global Recalls Portal.

IV. Build an **AI incidents tracker** to inform risk framework and policy [starting]
   - Develop an AI incidents and controversies tracker based on the media coverage of AI incidents (as a starting point) to calibrate the risk assessment work with evidence (with CSET, PAI, AIID)

V. Develop **Responsible Business Conduct guidance** for AI [in planning]
   - To help AI businesses proactively identify and address negative impacts that they might be causing or linked to through their operations and business relationships.
5 interrelated work strands:

I. Develop a **framework for classifying AI systems** *[completed]*

- To help policy-makers, regulators, legislators and others to assess the opportunities and risks that different types of AI systems present.

II. Develop a **risk assessment tool** to facilitate global interoperability *[starting]*

- Leveraging ongoing initiatives including in the EU, US, ISO, IEEE.
- Leveraging risk assessment work in other parts of the OECD.

III. Develop a common **AI incident reporting framework** *[in planning]*

- Framework for mandatory and voluntary reporting about AI incidents to ensure interoperability.
- Build on lessons learnt from OECD Global Recalls Portal.

IV. Build an **AI incidents tracker** to inform risk framework and policy *[starting]*

- Develop an AI incidents and controversies tracker based on the media coverage of AI incidents (as a starting point) to calibrate the risk assessment work with evidence (with CSET, PAI, AIID)

V. Develop **Responsible Business Conduct guidance** for AI *[in planning]*

- To help AI businesses proactively identify and address negative impacts that they might be causing or linked to through their operations and business relationships.
I. OECD framework for classifying AI systems

Why classify AI systems?
A variety of systems and policy implications
I. OECD framework for the classification of AI systems

Key dimensions characterise AI systems’ policy impact

Background
• Group of >80 experts
• Public consultation with >800 comments

www.oecd.ai/classification
I. OECD framework for the classification of AI systems

DATA & INPUT
- Provenance, collection, dynamic nature
- Rights and ‘identifiability’ (personal data, proprietary etc.)
- Appropriateness and quality

CONTEXT

ECONOMIC CONTEXT
- Industrial sector
- Business function & model
- Critical function
- Scale & maturity

PEOPLE & PLANET
- Users of the system
- Impacted stakeholders
- Optionality & redress
- Human rights, incl. privacy
- Well-being & environment
- Displacement potential

AI MODEL
- Model characteristics
- Model building (symbolic, machine learning, hybrid)
- Model inferencing / use

TASK & OUTPUT
- System task (recognise; personalise etc)
- System action (autonomy level)
- Combining tasks and action
- Core application areas (computer vision etc)
I. OECD framework for the classification of AI systems

AI System Lifecycle

DATA & INPUT
Actors include data collectors & processors
Collect & process data

ECONOMIC CONTEXT
Plan & design
Operate & monitor
Actors include system operators

PEOPLE & PLANET
Use or impacted by
Actors include end-users & stakeholders

AI MODEL
Build & use model
Verify & validate
Actors include developers & modellers

TASK & OUTPUT
Deploy
Actors include system integrators
Using the framework for health technology assessment

Clinical liability
- Current use
- Changes to data flows
- Local deployment factors
- Local performance

End-user training
- End-user well being
- End-user/patient autonomy
- Consent
- Communication strategies

Model reporting
- Training data reporting
- Performance across different groups
- Handling outlier data
- Harm

Output description
- Post deployment change management plan
- Benchmarking
- Oversight committees
- Autonomy
IV. Develop an AI incidents tracker

- Use AI incidents (incl. from media)
- Build evidence base of incidents to inform
  - AI risk assessments e.g. by calibrating risk estimations posed by certain AI systems
  - Regulatory choices e.g. by identifying high risk AI systems and threats (future risks)
  - AI incident reporting framework e.g. by identifying the relevant criteria to report

In collaboration with relevant players, e.g. the Responsible AI Collaborative (RAIC) and Georgetown’s CSET
IV. Proposed ‘Global AI media incidents tracker’

Step 1: Illustrative findings from 900+ manually identified news articles on AI incidents and controversies

*Caveat: significant sampling bias*

![Number of articles on AI incidents](chart)

IV. Proposed ‘Global AI media incidents tracker’

Step 2: automate AI incidents identification and classification into the criteria of the classification framework

Examples of automating AI incidents identification using news articles
V. Responsible Business Conduct (RBC) for Trustworthy AI

- Guidelines are well-established internationally (50 countries)
- Applied in numerous areas (e.g. minerals, agriculture and finance)
- Encompass companies of all sizes that operate internationally
- Include human rights (HR) due diligence, complementary to legislation
- Benefit from enforcement mechanism based on national contact points
  - A non-judicial grievance mechanism to facilitate resolution of disputes
  - Access to remedy on a global scale through platform for mediation and conciliation
  - Agreement reached in 42% of cases

Interesting option for accountability in a fast-moving area like AI
Catalogue of tools for trustworthy AI
Catalogue prototype

Catalogue of Tools for Trustworthy AI

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

Why we need a catalogue of tools for trustworthy AI

There are tools out there that help AI actors to build and deploy AI systems that are trustworthy. However, these tools can be hard to find and are ongoing AI policy discussions. This catalogue is a one-stop shop for AI actors to share approaches, mechanisms and practices to implement trustworthy AI in a comparable and easily accessible manner.

Tools by objective

- Accountability
- Privacy & data governance
- Fair & unbiased
- Reskill or upskill
- Transparent & explainable
- Help the planet
- Respect human rights
- Improve wellbeing
- Robust & secure

Tools by sector

- Agriculture
- Social & welfare issues
- Not applicable
- AI
- Industry & entrepreneurship
- Science & technology
- Investment
- Finance and insurance
- Trade
- Public governance
- Innovation
- Environment
- Tax
- Other
- Health
- Employment & labour

Submit a tool
Submit a use case
Submit a metric

Catalogue prototype


The OECD Policy Observatory offers a catalogue prototype of AI tools, allowing users to search and filter tools based on criteria such as approach, technical details, and safety. Examples include Foolbox, a Python library for adversarial attacks on machine learning models, and AIVerify AI Governance Testing Framework & Toolkit, which IMDA and PDPC have developed to assist industry in testing and demonstrating the deployment of responsible AI systems.
Foolbox

Organisation(s): Bernstein Center for Computational Neuroscience, International Max Planck Research School for Intelligent Systems, Tübingen AI Center, University of Tübingen

Foolbox is a Python library that lets you easily run adversarial attacks against machine learning models like deep neural networks. It is built on top of EagerPy and works natively with models in PyTorch, TensorFlow, and JAX.
Automated retrieval of technical tools from GitHub

- Training an AI system to identify tools and tag them automatically
Catalogue of tools for trustworthy AI: timeline and next steps

Q1 – Q2 2022
Produce mock-ups and test prototype

Q2-Q3 2022
Test, populate and improve the catalogue; refine vetting mechanisms; outreach for partnerships

Q4 2022
Launch catalogue; launch open calls for submissions

2023
Add sectoral focus; further develop content, use cases, tutorials
Thank you!

For more information please visit www.oecd.ai

e-mail: ai@oecd.org