Enabling effective AI policies: Launch of the OECD Framework for Classifying AI Systems

International Conference on AI in Work, Innovation, Productivity and Skills, 22 February 2022
Why classify AI systems?

A variety of systems and policy implications
OECD AI System Definition (OECD, 2019)

“An AI system, is a machine-based system that is capable of influencing the environment by producing an output (recommendations, predictions or decisions) for a given set of objectives.

i) perceives environments through data or input;
ii) abstracts these perceptions into models;
iii) uses the models to formulate options for outcomes.”
OECD Framework for Classifying AI systems: Key dimensions characterise AI systems’ policy impact
Linking the classification & AI system lifecycle actors

### Framework dimensions
- **People & Planet**
- **Economic Context**
- **Data & Input**
- **AI Model**
- **Task & Output**

### Actors include
- **End-users & stakeholders**
- **System operators**
- **Data collectors & processors**
- **Developers & modellers**
- **System integrators**

### Lifecycle stage
- **Use or are impact by**
- **Plan & design**
- **Operate & monitor**
- **Collect & process data**
- **Build & use**
- **Build & validate**
- **Deploy**
Uses of the OECD AI Classification Framework

**APPLICABILITY**: Most relevant to classifying specific AI applications, rather than generic AI systems

**GOAL**: Provide a baseline framework to help support and advance:

1. a common understanding of AI, and metrics.
2. structure registries or inventories of AI systems.
3. sector-specific frameworks, e.g. in healthcare (NICE).
4. risk assessment and incident reporting *(next steps)*.
5. risk management & work on accountability along the AI system lifecycle *(next steps)*.

**PROCESS:**

- Consensus of group of 60+ experts
- Testing & public consultation May-June 2021:
  - > 850 comments & survey responses => Adapted framework.

Sincere thanks for invaluable input to all who commented and tested the framework.
Each AI framework dimension has its own properties and attributes…

…and involves specific actors.
PEOPLE & PLANET

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

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

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

AI MODEL
- Model characteristics
- Model building
- Model evaluation (technique, 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)
ECONOMIC CONTEXT

- Industrial sector
- Business function & model
- Critical function
- Scale & maturity
DATA & INPUT

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

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

DATA & INPUT
- Optionality & redress
- Human rights, incl. privacy
- Well-being & environment
- Displacement potential

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

TASK & OUTPUT
- System task (recognise; personalise etc)
- System action (autonomy level)
- Combining tasks and action
- Core application areas (computer vision etc)
- System task (recognise; personalise etc)
- System action (autonomy level)
- Combining tasks and action
- Core application areas (computer vision etc)
Key conclusions from survey responses:

• The framework is best suited to **specific applications of AI systems** rather than to generic AI systems => the more specific the applications, the more consistent the survey responses.

• Respondents were better at classifying criteria in **People & Planet** and **Economic Context**. Classifying **Data & Input, AI Model, and Task & Output** often requires more technical information than is available publicly.
Example 1: Credit-scoring AI systems

Selected criteria:

- **System users** – Amateur (bank employee)
- **Optionality** – Cannot opt out
- **Human rights impact** – Yes
- **Sector of deployment** – Financial system (e.g., banking, insurance)
- **Critical function** – Critical function/activity (availability of financial services, inclusion)
- **Data collection** – Human (set of rules) and automated sources (e.g. profiles, loan payments)
- **Rights** – Mix of proprietary and public data
- **“Identifiability”** – Often personally identifiable data
- **Model building** – E.g., statistical/hybrid model; learns from provided data, augmented by human knowledge
- **Model evolution** – Can evolve during operation
- **System task** – Forecasting: uses past & existing behavior to predict future outcomes
- **Level of action autonomy** – Medium (human on-the-loop)
Selected criteria:

_Caveat: general purpose AI system, so nearly all responses depend on the specific application context! Medical advice, content filter, creative writing...

- **System users** – Primary users are amateur
- **Impacted stakeholders** – workers, consumers
- **Sector of deployment** – Information & communication
- **Critical function** – None
- **Data collection** – Human sources (text strings)
- **Rights** – Largely public data sources (some proprietary)
- **Model building** – Learn from provided data
- **Model evolution** – Evolution during operation
- **System task** – Goal-driven optimization, Reasoning with knowledge structures, interaction support, recognition, personalisation
- **Level of action autonomy** – Low autonomy [human action required e.g., to use generated text]
Using the framework to frame evidence standards for healthcare

1. **Scoping Review**

   - OECD classification framework was independently ranked as most complete system from a shortlist of 21 candidates when mapped against the 9 core domains of HTA (EUnetHTA)

2. **Questionnaire & Interview Study**

   - Highest rated by a global multistakeholder panel of experts in both a questionnaire and interview study

3. **Delphi Consensus Study**

4. **Public Consultation**

   - Framework now pivotal in developing evidence standards to underpin post-market evaluation in UK health sector
Using the framework for health technology assessment

The Alan Turing Institute

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

Input description

Model reporting
  Training data reporting
  Performance in different groups
  Handling outlier data
  Harm
  Autonomy

Output description
  Post deployment change management plan
  Benchmarking
  Oversight committees
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 the classification plus possible governance at the corporate, institution or AI systems level
  - 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.
Reminder: Risk categorisation of uses of AI in the draft AI Act of the EU

- **Unacceptable risk**
  - e.g. social scoring
  - Prohibited

- **High risk**
  - e.g. recruitment, medical devices
  - Not mutually exclusive
  - ‘Transparency’ risk
  - ‘Impersonation’ (bots)
  - Permitted subject to compliance with AI requirements and ex-ante conformity assessment

- **Minimal or no risk**
  - Permitted but subject to information/transparency obligations
  - Permitted with no restrictions
Details will be needed on European regulation/standardisation aspects …

• **Rules:**
  Mapping of uses of AI to these four categories (detailed rules, examples …)

• **Governance:**
  AI integration into / extension of frameworks for –
  - risk management
  - risk mitigation

• **Tools:**
  Assessment tools to operationalize mapping rules, due diligence, AI governance etc.
For more information visit [www.oecd.ai/classification](http://www.oecd.ai/classification)

email: ai@oecd.org