The OECD Framework for Classifying AI systems



Introducing the OECD Network of Experts on AI (ONE AI)



- Launched in February 2020
- 200+ Al experts from national governments, IGOs and the EC, business, civil society, academia, trade unions
- Facilitates collaboration between the OECD and other international initiatives on AI



Governments
 International Organisations
 European Commission
 Business
 Civil Society & Academia
 Technical community
 Trade unions





Three working groups and one task force to implement the **OECD AI Principles**

Working Group on Classifying Al systems

Values-based principles

Socio-economic & environmental impacts Human-centred values and fairness Transparency, explainability Robustness, security, safety Accountability

National Policies

Task Force on Al Compute

Investing in research Compute, data, technologies Enabling policy environment Jobs, skills, transitions International cooperation

WG on National AI policies

WG on

Tools for

Trustworthy

Α

Why classify Al systems? A variety of systems and policy implications





OECD AI System Definition (OECD, 2019)



"An **Al 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**.

It uses machine and/or human-based inputs/data to:

i) perceive environments;

ii) abstract these perceptions into models; and

iii) **use** the models to formulate options for **outcomes**.

Al systems are designed to operate with varying levels of autonomy."



What is the OECD Framework for the Classifying AI systems?



- A structure to differentiate Al systems & map implications for key policy areas framed by the OECD Al Principles
- A user friendly and simplified tool : 4 dimensions and 20 core criteria
- Linking technical aspects to policy considerations

The Framework has 4 dimensions & 20 core criteria



1. CONTEXT

- Industrial sector
- Business function
- Critical function
- Scale and maturity
- Users
- Impacted stakeholders, optionality, business model
- Human rights impact
- Well-being impact

Key actors include: system operators and end users

2. DATA AND INPUT

- Provenance, collection and dynamic nature
- Structure and format (structured etc.)
- Rights and 'identifiability' (personal, proprietary etc.)
- Appropriateness and quality

Key actors include: data collectors and processors

4. TASK AND OUTPUT

- Task of the system (recognition; personalisation etc.)
- Action of the system (autonomy level)
- Combining tasks and action
- Core application areas (computer vision etc.

Key actors include: system integrators

3. AI MODEL

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

Key actors include: developers and modellers

1. Context of the AI system

CONTEXT

Which industry?

For what business function?

How widely deployed?

What type of users?

Impacting whom?

With what degree of choice? What impact on human rights?

Performs a critical function?

How mature?

What business model?

What impact on well-being?

Key actors include: system operators and end users

Associated industry structure & regulation, <i>e.g.</i> , healthcare, transportation, public admin.
e.g., production; human resources management; sales;
Pilot stage, narrow, broad or wide deployment
AI experts or amateurs
e.g., consumers; employees; businesses; government; groups.
Whether users can or cannot opt out and / or challenge system output
System's potential on fundamental human rights (e.g; liberty; freedom of expression)
System performs critical function or is deployed in critical activity sector
Defined using NASA TRLs
For profit – advertising, subscription, other; non-profit
System's potential on well-being (e.g. on health, on job quality)
Core criteria

grey:

proposed as optional criteria that may be difficult to answer

Al System 2. DATA & INPUT 3. AI MODEL 4. TASK & OUTPUT

2. Data & input

DATA AND INPUT

Where does the data	Expert input, provided data, observed data, synthetic data, derived data
How is it collected?	By humans, by automated sensing devices, or both
How dynamic is it?	Static data, dynamic updated from time to time, dynamic real time
Is the data structured or not?	Unstructured, semi-structured, structured, complex structured
Are rights attached to the data / input?	Proprietary, public or personal data
What is dataset's scale?	Small, medium, large, very large
What is the data's format?	Standardised or non-standardised
Is the data identifiable?	Identified, pseudonymised, anonymised
Is the data "appropriate"?	Criteria to ensure that data are fit for the purpose
What is the data's quality?	Representative sample and sufficient sample size, completeness, data noise

Al System 2. DATA & INPUT 1. CONTEXT 4. TASK & OUTPUT 3. AI MODEL

grey: proposed as optional criteria that may be difficult to answer

3. Al Model



AI MODEL

What is the type of model used? How is the model trained / built?

Does the model evolve & how?

Is it discriminative? generative?

Does the model learn and how?

How is the model used?

Symbolic, statistical or hybrid

Human written-rules, supervised, semi-supervised, unsupervised learning;

No evolution, evolution through passive interaction, evolution through active interaction in machine learning models

Discriminative (i.e. predicting data labels by learning to distinguish between dataset classes) or generative model (i.e. explaining how the data was generated)

Centralised or federated learning

Deterministic (one outcome), probabilistic model (of distribution of outcomes) or both

Core criteria grey : proposed as optional criteria that may be difficult to answer

4. Task and output

TASK AND OUTPUT

What is the task performed by the system?

How autonomous are the system's actions and what role do humans play?

What is the labour displacement potential of the system?

Does the system belong to a core application area?

Does the system combine different tasks & actions ?

Recognition, event detection, forecasting, personalisation, interaction support, goal-driven optimisation, reasoning w. knowledge structures

No action autonomy, low, medium, high autonomy; Human support, human in-the-loop, on-loop, out-of-the loop

Potential for displacement through task automation

Human language technologies, computer vision, robotics

Autonomous system, control system, content generation

Core criteria grey: proposed as optional criteria that may be difficult to answer



Work in progress: Mapping systems' classification to risk



- Identified characteristics that signify a system is not low risk, including potential risks for human rights, building on Council of Europe's work
- Associated other characteristics with positive, negative or neutral impact on risk to obtain a preliminary cumulative effect

Next steps:

- Refine methodology and define output (*eg.* composite score)
- Test relevance & applicability

Al system characteristics (by dimension)

	risk	risk
1) CONTEXT		
Industrial sector	↑ or ↓	
Business function	↑ or ↓	
Impacts critical functions / activities		
AI system is in a critical sector or infrastructure	↑	
AI system performs a critical function independent from its sector	↑	Χ
Breadth of deployment		
A pilot project		
Narrow deployment (e.g. one company in one country)	<u>↑</u> or ↓	
Broad deployment (e.g. one sector)	^	
Widespread deployment (e.g. across countries and sectors	^	
Al system maturity		
TRL 1 to 3	^	
TRL 4 to 7	<u>↑</u> or ↓	
TRL 8 to 9	↓	
Users of AI system		
Amateur	<u>^</u>	
Practitioner who is not an AI expert	↑ or ¥	
Practitioner who is an AI expert or system developer:	↑ or ↓	
Al system maturity	↑ or ↓	
Impacted stakeholders		
Consumers	^	
Workers / employees	<u>↑</u>	
Business	↑ or ¥	
Government agencies / regulators	<u>^</u>	
Specific communities	↑ or ¥	
Children or other vulnerable or marginalised groups	$\mathbf{\hat{T}}$	
Optionality		
Users cannot opt out of using the AI system		
Users can correct or contest AI output	↑ or ♥	
Users can opt-out of using the system	•	
For-profit use non-profit use or public sector use		

Cumulative

effect on

Not

low

Direct

Direct

From classification to risk assessment (1)

Non-profit use (outside public sector)	↑ or ↓	
Public sector use	•	
Other	↑ or ↓	
and immediate risks of violating human rights or fundamental values (on	nly considering negative imp	acts)
Life and physical and mental integrity	1	Х
Liberty and security	^	Х
Fair trial; no punishment without law; effective remedy	^	Х
Privacy and family life	^	
Freedom of thought, conscience and religion	^	Х
Freedom of expression; assembly and association	^	х
Non-discrimination	^	
Protection of property and peaceful enjoyment of possessions	^	
Right to education	^	Х
Right to democracy and free elections	^	Х
Human autonomy	^	
Human dignity	^	
Other (detail)	^	
and immediate risks to individuals' well-being (only considering negative	e impacts)	
Health (including mental health)	^	Х
Housing	↑	Х
Income and wealth	^	
Work and job quality	↑	
Environment quality	↑	
Social connections	^	
Civic engagement	^	
Education	1	
Subjective well-being	<u>^</u>	
Work-life balance	1	
	1	.3

Note: items marked " \uparrow or \checkmark " are to be assessed depending on the AI system usage and outcomes.

Al system characteristics (by dimension)	Cumulative effect on risk	Not Iow risk
I) CONTEXT	Hor	non
Industrial sector		
Business function		
Impacts critical functions / activities	🛧 or 🕹	
Al system is in a critical sector or infrastructure		
Al system performs a critical function independent from its sector	🛧 or 🖖	
Breadth of deployment		
A pilot project		1.0
Narrow deployment (e.g. one company in one country)	\bullet	i c
Broad deployment (e.g. one sector)		Y
Widespread deployment (e.g. across countries and sectors		Λ
Al system maturity		
TRL 1 to 3	1	
TRL 4 to 7	Annuk	
TRL 8 to 9	↑ or ♥	
Users of Al system		
Amateur	$\mathbf{\uparrow}$	
Practitioner who is not an AI expert	î or ↓	
Practitioner who is an AI expert or system developer:	🋧 or 🗸	
Al system maturity	î di 🗸 🔨	
Im pacted stakeholders		Direct a
Consumers		
Workers / employees		
Business	🛧 or ₩	
Government agencies / regulators	\uparrow	
Specific communities	🛧 or ₩	
Children or other vulnerable or marginalised groups		
Optionality		
Users cannot opt out of using the AI system	^	
Users can correct or contest AI output	î de vere de la constante de	
Users can opt-out of using the system	↓	
Ear and States and an addition on analytic constant and		

From classification to risk assessment (1)

Public sector use	Non-profit use (outside public sector)	T or T	
Other 			
Ide and physical and mental integrity Image: Construction of the physical and mental integrity Image: Construction of the physical and mental integrity Liberty and security Image: Construction of the physical and mental integrity Image: Construction of the physical and mental integrity Image: Construction of the physical and mental integrity Privacy and family life Image: Construction of the physical and mental integrity Image: Construction of the physical and mental integrity Image: Construction of the physical and mental integrity Privacy and family life Image: Construction of the physical and mental integrity		小 or Ψ	
Life and physical and mental integrity Liberty and security Fair thal, no punishment without law, effective remedy Privacy and family life Freedom of thought, conscience and religion Freedom of expression, assembly and association Non-discrimination Protection of property and peaceful enjoyment of possessions Right to education Right to democracy and free elections Human autonomy Human dignity Other (detail) Mork (detail) Health (including mental health) Housing Income and wealth Work and job quality Environment quality Social connections Civic engagement Education Subjective well-being Work-life balance	nd immediate risks of violating human rights or fundamental values (only considering	g negative imp	acts)
Liberty and secunty Image: Security Fair thail, no punishment without law, effective remedy Image: Security Privacy and family life Image: Security Freedom of thought, conscience and religion Image: Security Freedom of expression, assembly and association Image: Security Non-discrimination Image: Security Protection of property and peaceful enjoyment of possessions Image: Security Right to education Image: Security Right to education Image: Security Human autonomy Image: Security Human dignity Image: Security Other (detail) Image: Security Health (including mental health) Image: Security Housing Image: Security Income and wealth Image: Security Vork and job qualty Image: Security Environment quality Image: Security Social connections Image: Security Givic engagement Image: Security Education Image: Security Subjective well-being Image: Security	Life and physical and mental integrity		Х
Fair trial, no punishment without law; effective remedy Privacy and family life Freedom of thought, conscience and religion X Freedom of expression, assembly and association X Non-discrimination Protection of property and peaceful enjoyment of possessions Right to education X Right to democracy and free electons X Human autonomy X Human dignity X Other (detail) X X Health (including mental health) X X Housing X X X X X More and job quality X X X X More and wealth X X X X X X X<	Liberty and security		
Privacy and family life Freedom of thought, conscience and religion Teedom of expression, assembly and association Non-discrimination Protection of property and peaceful enjoyment of possessions Right to education Right to democracy and free elections Tere (detail) Income and weath Work and job quality Environment quality Social connections Civic engagement Education Subjective well-being Work-life balance	Fair trial, no punishment without law, effective remedy		
Freedom of thought, conscience and religion Freedom of expression, assembly and association Non-discrimination Protection of property and peaceful enjoyment of possessions Right to education X Right to democracy and free elections X Human autonomy X Human dignity Y Other (detail) Y Income and weath Y Social connections Y Social connections Y Subjective well-being Y Y Y Y Subjective well-being Y Y 			
Freedom of expression, assembly and association Non-discrimination Protection of property and peaceful enjoyment of possessions Right to education Right to democracy and free elections Ruman autonomy Ruman dignity Other (detail) Health (including mental health) Housing Income and wealth Work and job quality Environment quality Social connections Civic engagement Education Subjective well-being	Freedom of thought, conscience and religion		
Non-discrimination Image: Constraint of possessions Protection of property and peaceful enjoyment of possessions Image: Constraint of possessions Right to education Image: Constraint of possessions Right to democracy and free elections Image: Constraint of possessions Human autonomy Image: Constraint of possessions Human dignity Image: Constraint of possessions Other (detail) Image: Constraint of possessions nd immediate risks to individuals' well-being (only considering negative impacts) Image: Constraint of possessions Health (including mental health) Image: Constraint of possessions Image: Constraint of possessions Income and wealth Image: Constraint of possessions Image: Constraint of possessions Image: Constraint of possessions Cwic engagement Image: Constraint of possessions Image: Constraint of possessions Image: Constraint of possessions Subjective well-being Image: Constraint of possessions Image: Constraint of possessions Image: Constraint of possessions Work-life balance Image: Constraint of possessions Image: Constraint of possessions Image: Constraint of possessions Work-life balance Image: Constraint of possessions Image: Consestraintof possessions Image: Constraint	Freedom of expression, assembly and association		
Protection of property and peaceful enjoyment of possessions Right to education Right to democracy and free elections Human autonomy Human dignity Other (detail) Right including mental health Housing Right and job quality Environment quality Social connections Crvic engagement Education Subjective well-being Work-life balance 	Non-discrimination		
Right to education Image: Constraint of the elections Right to democracy and free elections Image: Constraint of the elections Human autonomy Image: Constraint of the elections Human dignity Image: Constraint of the elections Other (detail) Image: Constraint of the elections Health (including mental health) Image: Constraint of the elections Housing Image: Constraint of the elections Income and wealth Image: Constraint of the elections Work and job qualty Image: Constraint of the elections Crivic engagement Image: Constraint of the elections Education Image: Constraint of the elections Subjective well-being Image: Constraint of the elections Work-life balance Image: Constraint of the elections	Protection of property and peaceful enjoyment of possessions		
Right to democracy and free elections Human autonomy Human dignity Other (detail) Other (detail) The alth (including mental health) Housing Income and wealth Work and job quality Social connections Civic engagement Educaton Subjective well-being Work-life balance Model and the balance 			
Human autonomy Image: Constraint of the second			
Human dignity Image: Construction of the second of the			
Other (detail) nd immediate risks to individuals' well-being (only considering negative impacts) Health (including mental health) 			
Health (including mental health)			
Health (including mental health) Image: Constraint of the second sec	nd immediate risks to individuals' well-being (only considering negative impacts)		
Housing Income and wealth			
Income and wealthWork and job qualityEnvironment qualitySocial connectionsCivic engagementEducationSubjective well-beingWork-life balance			
Work and job quality	Income and wealth		
Environment quality	Work and job quality		
Social connections Civic engagement Civic engagement			
Civic engagement function Education † Subjective well-being † Work-life balance †			
Education Subjective well-being Work-life balance	Civic engagement		
Subjective well-being Work-life balance			
Work-life balance	Subjective well-being		
	Work-life balance		

Note: items marked " \uparrow or \checkmark " are to be assessed depending on the AI system usage and outcomes.

14

From classification to risk assessment (2, 3, 4)

2) DATA AND INPUT

Provenance of data and input	î or ♦
Detection and collection of data and input	↑ or ↓
Dynamic nature of data	
Static data	↓
Dynamic data updated from time-to-time	↑ or ↓
Dynamic real-time data	^
Scale	↑ or ↓
Structure of data and input	↑ or ↓
Format of data and metadata	
Standardised data format	↑ or ↓
Non-standardised data format	^
Standardised dataset metadata	↑ or ↓
Non-standardised dataset metadata	^
Rights associated with data and input	
Proprietary data	^
Public data	↑ or ↓
Personal data	^
Identifiability of personal data	
Identified data	▲
Pseudonymised data	↑ or ↓
Unlinked pseudonymised data	↓
Anonymised data	↓
Aggregated data	↓
Data quality and appropriateness	
appropriateness of data for a particular problem	↓
(high) sample representativeness	↓
adequate sample size	↓
(high) completeness and coherence of sample	↓
(low) data noise	J.

Note: items marked " \uparrow or \checkmark " are to be assessed depending on the AI system usage and outcomes.



3) AI MODEL	
I model characteristics	
(High) transparency and explainability	↓
(High) safety, security, robustness	↓
(High) reproducibility	↓
Evolution during operation	^
Evolution through uncontrolled learning	<u>^</u>
Privacy-preserving properties, e.g. federated learning	↓
4) TASK AND OUTPUT	
ask of the system	
Recognition	↑ or ↓
Event detection	↑ or ↓
Forecasting	<u></u> ↑ or
Personalisation	^
Interaction support	

 Interaction support
 ↑

 Goal-driven optimisation
 ↑ or ↓

 Reasoning with knowledge structures
 ↑ or ↓

 Action autonomy level
 ↑

 High action autonomy
 ↑

 Medium action autonomy
 ↑

 Low action autonomy
 ↑ or ↓

 No autonomy
 ↓

 Displacement potential
 ↑

 High displacement potential
 ↑ or ↓

15

Example 1: Al-powered hiring system

2. DATA AND INPUT

1. CONTEXT

Sector: Administrative and support services **Business function**: Human resources **Critical function**: No Scale: Narrow System users: Amateur (job candidate and HR) Impacted stakeholders: applicants / employees Human rights impact: YES Well-being impact: YES

Data provenance: Provided by candidate; Observed by algorithms; Derived Collection: Collected by automated tools Dynamism: Static; dynamic; real-time Structure : Unstructured data Rights & Identifiability: Proprietary personal data; identified Appropriateness & quality: Unknown

4. TASK AND OUTPUT

Task: Personalization, interaction support, recognition
Autonomy level: Medium – provides recommendation for human decision execution
Composite system: Yes
Belongs to core application area: TBD



3. AI MODEL

Model type: Hybrid Model building / training: Semi-supervised Discriminative model Model inferencing / use: Probabilistic

Example 2: Credit scoring system



3. AI MODEL

Model type: Hybrid (statistical & symbolic)

Model building / training: Acquisition from data, Augmented by human-encoded knowledge

Learning mode: Central learning

Model inferencing / use: Deterministic

1. CONTEXT

Sector: Financial and insurance activities **Business function**: Sales & customer service **Critical function**: YES Scale: Broad System users: Non-Al expert (bank employee) Impacted stakeholders: consumers **Optionality**: None Human rights impact: YES Well-being impact: YES

2. DATA AND INPUT

Provenance: expert rules; observed by algorithms (eg. payments history); provided by consumer (eg, name); derived (eg. credit score)
Collection: By humans & by sensing devices
Dynamism: Static (eg, gender), dynamic and realtime (eg. payments)
Structure & Format: structured & standardised
Rights: Personal, proprietary, identified data
Appropriate for purpose ; Quality unknown

4. TASK AND OUTPUT

Task: Forecasting & reasoning with knowledge structures

Autonomy level: Low

Composite system: YES (combines tasks and actions) **Belongs to core application area**: TBD

Example 3: Manufacturing plant management system

1. CONTEXT

Sector: Construction Business function: Sales, logistics, HR, monitoring ... Critical function: No Scale: Narrow Maturity: TRL 9 System users: Amateurs, non-expert and experts Impacted stakeholders: Workers, consumers, firms **Optionality**: Variable Human rights impact: YES Well-being impact: YES

2. DATA AND INPUT

Provenance: all (expert rules; observed by algorithms; provided by consumer; derived; synthetic)

Collection: by humans & by sensing devices **Dynamism:** Static (human knowledge), dynamic real-time (from production lines)

Structure & Format: all types of structures; standardized and un-standardized format **Rights:** Proprietary

Appropriate for purpose **; Representative** and appropriate, noise /missing values/outliers known

4. TASK AND OUTPUT

Task: All (forecasting; recognition; reasoning...)
Autonomy level: Medium
Displacement potential: High
Composite system: Yes
Belongs to core application area: Human language technologies (as well as process planning and

optimisation and Internet-of-things)

OECD.AI Policy Observatory

3. AI MODEL

Model type: Hybrid (statistical & symbolic)

Model building / training: Acquisition from data, Augmented by humanencoded knowledge

Learning mode: Federated

Data interaction: Evolution during operation through active and passive interaction

Model inferencing / use: both deterministic & probabilistic

How to participate in the consultation?



~

oecd.ai/p/classification

1. Comment on report :

provide comments and suggestions directly on the written version of the Framework.

2. Take our online survey to test the Framework on an Al system of your choice.





The OECD AI Systems Classification Framework

This survey is currently not active. You will not be able to save your responses.

FIRST DIMENSION: CONTEXT

The context or socio-economic context in which an AI system operates has key consequences for policy. Core characteristics (marked in red with a *) of the "Context" dimension include the sector in which an AI system is deployed, its deployment impact and scale, its effects on human rights and well-being, and its critical (or non-critical) nature.

Note: Only questions relating to core characteristics are compulsory. While we encourage filling as many questions as possibles, other (non-core) questions are optional and require a higher degree of technical knowledge.

* 1. Select the sector in which the AI system is deployed:

O More detail about the industries included in each ISIC (rev 4) sectors listed below can be found here

Please choose...

* 2. Select the business function performed by the AI system:

Please choose...

Public consultation on the Framework *Key questions*



- 1. Should **core and and non-core criteria** be distinguished? I.e. should there be a core classification framework for less-technical audiences + additional considerations for more technical / informed users?
- 2. Which **characteristics** should be **core criteria** and which 'optional'?
- 3. Can AI systems be classified **consistently & reliably** with the core criteria?
- 4. Which criteria are useful for a more detailed, technically-oriented framework?
- 5. Should there be **industry or application domain specific criteria** and classifications, e.g. depending on context?

Timeline for the consultation



oecd.ai/p/classification

- 1. Timeline: 20 May 2021 30 June 2021
- 2. Comments summarised and published: summer 2021
- 3. Planned launch for the Framework: fall 2021